Spatial epidemiology of cancer: a review of data sources, methods and risk factors

Rita Roquette,1,2 Marco Painho,1 Baltazar Nunes2,3
1Nova IMS Information Management School, Lisbon; 2National Health Institute Doutor Ricardo Jorge, Lisbon; 3Escola Nacional de Saúde Pública, Universidade Nova de Lisboa, Lisbon, Portugal

Abstract
Cancer is a major concern among chronic diseases today. Spatial epidemiology plays a relevant role in this matter and we present here a review of this subject, including a discussion of the literature in terms of the level of geographic data aggregation, risk factors and methods used to analyse the spatial distribution of patterns and spatial clusters. For this purpose, we performed a web-search in the Pubmed and Web of Science databases including studies published between 1979 and 2015. We found 180 papers from 63 journals and noted that spatial epidemiology of cancer has been addressed with more emphasis during the last decade with research based on data mostly extracted from cancer registries and official mortality statistics. In general, the research questions present in the reviewed papers can be classified into three different sets: i) analysis of spatial distribution of cancer and/or its temporal evolution; ii) risk factors; iii) development of data analysis methods and/or evaluation of results obtained from application of existing methods. This review is expected to help promote research in this area through the identification of relevant knowledge gaps. Cancer’s spatial epidemiology represents an important concern, mainly for public health policies design aimed to minimise the impact of chronic disease in specific populations.

Introduction
Given the relevance of spatial epidemiology in health research and the emphasis of cancer among chronic diseases, as well as the growing amount of studies in this area, it is important to know what the literature says about spatial epidemiology of cancer as well as provide its structured description. According to the World Health Organization (WHO), cancer is a leading cause of death in the world (WHO, 2015). It is also the cause of various morbidities and co-morbidities and can be responsible for loss of years of life years as well as loss of years without disability. Considering the aging population, it is predicted that the number of new cases of cancer will increase by more than 12% over the next decade in the European Union (EU) (DGS, 2013). The fight against cancer is a major challenge in public health. This challenge is due in part to the inequalities in terms of incidence, mortality, and survival. Therefore, a multidisciplinary approach is needed (Bastos et al., 2010). Among the various fields that can contribute to the development of knowledge about this disease, spatial epidemiology plays an important role. It can promote the understanding of spatial and temporal distribution patterns, helping to better identify the risk factors that influence them.

Three types of approach can be established in spatial epidemiology: i) mapping; ii) geographic correlation; and iii) clustering (Elliott and Wartenberg, 2004). Mapping or map design regarding health and disease situations is the most often mentioned and used of these three approaches. Further, geographic correlation studies have the goal to spatially compare the health with several types of factors such as environmental, economic, social, demographic or lifestyle (Elliott and Wartenberg, 2004). They can also give clues to the investigation of disease causes (Wakefield, 2004). Finally, concerning the third approach, clustering could be the most relevant from an epidemiologic point of view (Clarke et al., 1996). Cluster can be defined as an unusual agglomeration of high or low occurrence of a phenomena (Lawson, 2010).

A search of literature reviews about spatial epidemiology in the Web of Science (Reuteurs, 2016) revealed three main articles, although two of them do not specifically analyse studies related to cancer. Auchincloss and colleagues (2012), in A Review of Spatial Methods in Epidemiology, 2000-2010, refer the growing number of articles in the spatial epidemiology field based on articles pub-

This article presents a literature review about cancer’s spatial epidemiology. In particular, considering that literature review is a generic term, our article presents a systematized review (Grant and Booth, 2009). Literature is discussed in terms of the level of geographic data aggregation, risk factors, and methods applied to analyse the spatial distribution of patterns and spatial clusters. The innovation of this study concerns the use of a different approach compared to the reviews described above, i.e. it considers cancer specifically and uses a longer period of years of publication to describe the evolution of the volume of published papers and the main subjects covered and also points out gaps of knowledge.

**Materials and Methods**

We performed a systematized review (Grant and Booth, 2009) using the databases Pubmed (accessed at 20th July 2016) and Web of Science (accessed at 28th April 2016). We considered all papers published from 1979 to 2015. Since the search fields available in both databases are different, the search on each one of them was slightly different too. For example, we searched *Title* and *Title/Abstract* in Pubmed and *Title* and *Topic* in the Web of Science. Thus, the search covered all papers published until the end of 2015 that included the following terms: a) ((((cancer *[Title]*) OR (neoplasm *[Title]*) AND epidemiology [Title/Abstract]) AND spati*[Title/Abstract]) AND Geographic*[Title/Abstract]) AND cluster [Title/Abstract]) OR (((((cancer [Title]) OR (neoplasm [Title])) AND epidemiology [Title/Abstract]) AND spati*[Title/Abstract]) AND Geographic*[Title/Abstract]) AND model [Title/Abstract]) OR (((((cancer [Title]) OR (neoplasm [Title])) AND epidemiology [Title/Abstract]) AND spati*[Title/Abstract]) AND Geographic*[Title/Abstract]) AND model [Title/Abstract]) in the Pubmed database or; b) ((TI=(cancer) OR TI=(neoplasm)) AND TS=(spati*) AND TS=(epidemiology) AND TS=(geographic*) AND TS=(cluster)) OR ((TI=(cancer) OR TI=(neoplasm)) AND

![Figure 1. Spatial epidemiology review of cancer: search scheme followed.](image)

![Figure 2. Annual distribution of the published articles selected in the review.](image)
TS=(spati*) AND TS=(epidemiology) AND TS=(geographic*) AND TS=(distribution)) OR ((TI=(cancer) OR TI=(neoplasm)) AND TS=(spati*) AND TS=(epidemiology) AND TS=(geographic*) AND TS=(model)) OR ((TI=(cancer) OR TI=(neoplasm)) AND TS=(spati*) AND TS=(distribution) AND TS=(geographic*) AND TS=(model)), in the Web of Science database.

After article selection, we analysed the scientific areas of the journals in order to refine the search criteria and select the papers more related to the subject addressed. This was performed considering the subject categories associated with each journal at the Scimago website (Scimago, 2007). From the diversity of categories, we selected those considered directly related to the subject of our search, namely Public health, environmental and occupational health; Geography, planning and development; Epidemiology; Oncology; Health (social science); and Health toxicology and mutagenesis. The articles’ search scheme, summarising all search steps, is represented in a PRISMA flow diagram (Moher et al., 2010) presented in Figure 1.

The spatial epidemiology of cancer can be very wide-ranging and covered in many different papers. Therefore, as we didn’t want to refine our search in terms of publication data, we had to be more restrictive in the choice of search terms. Nevertheless, we tried to ensure that these terms covered a variety of thematic perspectives. The literature was analysed in terms of publication date, keyword, cancer site, data source, observation unit, study objective, risk factor and applied method, as referred in the following sections. The key features of the papers were summarised and described in tables and graphs. Quantitative synthesis was also performed, using descriptive statistics.

Results

General

Our search resulted in a selection of 180 articles from 63 journals. As shown in Figure 2, few articles were published in the early years of the period investigated (1979-2015). The growth in the number of articles started in 2002, grew irregularly until the peak in 2012, with 18 articles published, after which the boom subsided. In terms of the journals with published articles about spatial cancer distribution, Table 1 lists those with more than five articles published.

Key words

Only about 56% of the articles contained keywords. This seems to be related to the journals’ publication rules. Although we expect that search options influence the resulting keywords, it is interesting to analyse which ones are the most popular. For this propose, was constructed a tag cloud graph [using TagCrowd software (Steinbock, 2006)], which allowed to present the most cited keywords, at a maximum of 50 keywords. The representation of words was made in light of their frequency (Figure 3).

The most frequent keywords were: i) cancer, cited in 14% of the articles; ii) disease mapping, cited in 13%; iii) GIS and epidemiology, each one cited in 12%; iv) breast cancer, spatial analysis and GIS, each one being cited in 11%; v) cancer incidence, lung cancer and prostate cancer, each one presented in 9%; vi) spatial epidemiology in 8%; and, finally, vii) cluster analysis and colorectal cancer, each cited in 6% of the articles. We considered the keywords individually as they were referred to by the authors. However, if the words geographic information system and geographical information system and GIS were considered together, this keyword becomes the most cited, being present in a total of 26% of the articles.

Table 1. Distribution of articles according to the journal of publication.

<table>
<thead>
<tr>
<th>Journal</th>
<th>Articles (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Journal of Health Geographics</td>
<td>26</td>
</tr>
<tr>
<td>BMC Cancer</td>
<td>10</td>
</tr>
<tr>
<td>Cancer Epidemiology</td>
<td>9</td>
</tr>
<tr>
<td>Cancer Causes and Control</td>
<td>9</td>
</tr>
<tr>
<td>Statistics in Medicine</td>
<td>8</td>
</tr>
<tr>
<td>Revue D’Epidemiologie et De Santé Publique</td>
<td>8</td>
</tr>
<tr>
<td>Health and Place</td>
<td>6</td>
</tr>
<tr>
<td>Environmental Health</td>
<td>6</td>
</tr>
<tr>
<td>Spatial and Spatio-Temporal Epidemiology</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 3. Most cited keywords in the articles selected in the review.
Cancer by site

A total of 28 cancer sites are cited in the reviewed articles. About 59% of the articles refer only one cancer site, and 38% refer to two or more sites. The remain articles do not refer to a particular cancer site, considering cancer as a whole, or they indicate various types of cancer but did not mentioned exactly which (Bhowmick et al., 2008; Hendryx et al., 2012; Ruktanonchai et al., 2014). Table 2 summarises the frequency of all cancers analysed, classified according to the 9th revision of International Statistical Classification of Diseases and Related Health Problems (ICD). It provides relative frequencies since an article can address more than one cancer site that makes the interpretation of the number of articles per cancer site difficult.

The cancer sites most frequently studied are malignant neoplasm of bone, connective tissue, skin and breast (especially due breast cancer research) and malignant neoplasm of the genitourinary organs (major part related to prostate cancer analysis).

Data sources used

Data analysis that can be performed on cancer’s spatial epidemiology depends first and foremost on the data disaggregation. From the selected articles, six (Bhowmick et al., 2008; Jia et al., 2014; Klassen and Platz, 2006; Lower, 1982; Tuyns and Repetto, 1979; Wan et al., 2012) are essentially theoretical and/or methodological, or did not use cancer data. The others are listed in Table 3 and classified according to data aggregation and data sources.

Most of the articles (about 70%) are based on individual data (Figure 4). Some of these articles used data from both registries and databases from projects or programs (Chien et al., 2013a; Gallagher et al., 2010), while others used individual data and aggregated data (Kulldorff et al., 2006). It should be noticed that a high proportion of the articles used cancer incidence data, i.e. all the articles based on aggregated data collected in databases developed within projects or programs and almost all articles representing individual data.

Observation unit used in the articles

As mentioned before, more than half of the articles used data sources of individually disaggregated cancer data. However, in many of these cases, the data were aggregated into areas prior to analysis. Thus, about 74% of the articles have analysed cancer data aggregated by geographic area.

Objectives of the studies presented in the articles

It was found a great variability in the articles objectives. However, independently of their aim or approach, all of them considered somehow geography as an issue. We identified three types of research approaches in the articles (listed and summarised in Table 4). The studies that analysed spatial distribution and/or temporal evolution of disease were applied to concrete and diversified geographical areas. Therefore, we decided not to present a detailed description of their findings. The other two groups of research questions are detailed in the next sections.

Risk factors

The association between cancer morbidity or cancer mortality and possible risk factors is discussed in many of the articles reviewed here (61%), although most of them are not conclusive. The majority provide information regarding the factors that may promote the occurrence of disease but also mention the need of fur-
In their research to confirm the results. In order to synthesise the factors considered in each article, we classified the factors covered into four groups: demographics and socioeconomics; environmental issues; individual behaviour; and physiological and genetic topics. Results using this classification have been reported in the form of a so-called Venn diagram (Oliverus, 2007-2015). Figure 5 shows that demographic and socioeconomic factors together with environmental factors were those most considered for analysis. Further, physiological and genetic factors were analysed more times than individual behaviour. Only very few articles included factors from all the groups.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Source</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Databases from projects or programmes</td>
<td>Death certificates</td>
<td>Chien et al. (2013b), Wei et al. (2012)</td>
</tr>
<tr>
<td>Databases from projects or programmes</td>
<td>Individual data</td>
<td>Asmarian et al. (2013), Boffetta et al. (2006), Christakos and Lal (1997), Mahaki et al. (2011), Mandal et al. (2009), Schootman et al and Sun et al. (2004), Schootman et al. (2011), Senkowski et al. (2008), Short et al. (2002), Toledano et al. (2001), Tsai et al. (2011), Van Leeuwen et al. (1999), Yu et al. (2013)</td>
</tr>
</tbody>
</table>
**Discussion**

The results based on the 180 papers dealing with spatial epidemiology of cancer show that, there was a large increase in the number of papers published in the last decade. This could possibly be due to the increased and now widespread use of computers as well as the generalisation of GIS adoption. GIS appeared in Canada in the 1960s (Tomlinson, 1998) with the aim of acquiring, storing, and processing geographical and alphanumeric information. It allowed the visualisation of both data and results based on such information. GIS usage has become popular in research since 1980s (Bhopal, 2008). Nowadays, we can realise that GIS has a great potential in public health and epidemiology for decision making and research (Clarke et al., 1996).

With regard to the site of the cancers considered in the analysed papers, those most commonly referred to were those known to have the highest incidence rates globally. Some of the most popular keywords that were found in the different papers were: disease mapping; cancer sites (not all sites given but commonly lung and breast); and incidence. This supports the idea that our search criteria seem sufficiently comprehensive to select items with different approaches to the topic under search, using different statistical methods (e.g., mapping, clusters) and addressing several cancer sites with the focus on different epidemiological measures (e.g., incidence).

Cancer incidence data were used in most of the papers reviewed here. The use of incidence rates can be preferable to mortality data from official national statistics since the former can i) provide information on anatomical and histological characteristics of cancer; ii) better describe the extent of the problem of disease in populations; iii) and facilitate comparison of data between countries (Christakos and Lai, 1997). In addition, the survival rate of one cancer site may vary according to geographical area (due to the medical conditions), which may hamper the geographical comparison of mortality data (Horner and Chirikos, 1987).

The analysis of epidemiological study designs could be a very interesting matter in the scope of this study. Nevertheless, a large number of the analysed papers did not clearly describe the epidemiological study design. For that reason, it was not possible to present this information consistently.

As mentioned before, more than a half of the articles reviewed used data sources of individually disaggregated cancer data, which were, in many of the studies, aggregated into areas prior to analysis. Thus, in 74% of the articles, the cancer analysis unit consisted of data aggregated by geographic area. In spatial epidemiology research, the degree of the data’s geographical aggregation is a very important issue. Both the use of disaggregated data at the individual level or at large geographical scales (as, for instance when zip codes are used) and the use of aggregated data at small scales have positive and negative aspects. On the one hand, the positive aspects of using disaggregated data are related to the greater variety of possible analytical approaches. There are some analysis methods that are only applicable to individual data (see, for instance, Timander and McLafferty, 1998). On the other, the major problem of using highly disaggregated data is the difficulty of ensuring data confidentiality and anonymity of individuals (Goovaerts, 2005). Also, the use of highly disaggregated geographical data implies a small number of occurrences of a given disease in each area, which makes it difficult to obtain precise statistical

---

**Table 5. Common spatial issues covered by research papers and corresponding methods.**

<table>
<thead>
<tr>
<th>Issue</th>
<th>Method applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoothed rate estimation</td>
<td>Rate calculation in areas with small populations can lead to resultant unstable rates, which can often be mitigated by calculating smoothed rates. The more often referred models in the articles reviewed were Bayesian (both empirical and fully Bayesian models) and Kriging. Among all the models Besag, Molie and York stood out.</td>
</tr>
<tr>
<td>Cluster analysis</td>
<td>In cluster analysis the most used methods were spatial scan statistics, and Moran’s I (both global and local)</td>
</tr>
<tr>
<td>Factor studies</td>
<td>Methods such as hierarchical modelling, multilevel modelling, logistic regression or geographical weighted regression were mentioned in studies of factors or in environmental context analysis</td>
</tr>
</tbody>
</table>

*See, for instance, Lopez-Abente et al. (2014) and the webpage: https://www.math.stevens.edu/~brunhilde/da/doc/latest/besag.pdf*
values (Chiang et al., 2010; Fairley et al., 2008; Goovaerts, 2006b). This problem, called the small numbers problem (Goovaerts, 2005; Shi, 2009), is further enhanced when the diseases investigated are rare (Thompson et al., 2007) and/or the populations of the geographical units under analysis are small (Goovaerts, 2006a; Short et al., 2002).

The difficulties of confidentiality and reliability of highly disaggregated data (Chien et al., 2013b) are more commonly addressed by aggregating the data at small scales or over of long time (several years). One of the benefits of aggregated data at small scales is the mitigation or absence of the small numbers problem. However, the larger the data aggregation, the greater the probability of agglomerations with high or low values occupying only part of the area under analysis resulting in hidden information or average dilution of the whole geographic area under investigation (Fang et al., 2004). It is not possible to state, precisely, which degree of data disaggregation is the most appropriate for an analysis in spatial epidemiology and this remains, in fact, a controversial topic. This controversy extends to the question of the stability of calculated statistical measures (like incidence rate). There are studies in which the data are aggregated in order to reduce the uncertainty associated with the analysis results (Huang et al., 2010). However, some authors argue that the spatial pattern of aggregated data could result from aggregation methods rather than data themselves (Krewski et al., 2005).

Regardless the degree of aggregation, the performance of an aggregated data analysis should take into account some concerns, among which the following stand out. First, in combined analysis of geographically aggregated data, difficulties may arise when they are not grouped according to the same geographical boundaries (Blackley et al., 2012; Goovaerts, 2006a); second, analysis results of aggregated data should be considered true only at their scale of aggregation and should not be extrapolated to other aggregation or disaggregation levels (Fortunato et al., 2011) since inconsistencies in results obtained at different scales may arise (Goovaerts and Xiao, 2011, 2012); third, the spatial patterns obtained based on aggregated data can result from the level of aggregation chosen and not from the distribution of the phenomenon under review itself (Krewski et al., 2005); and fourth, data are often aggregated into geographical areas defined for political or administrative reasons (Gregorio et al., 2006), which may not always be the most appropriate for undertaking a particular study (Goovaerts, 2006a). If the areas’ aggregation criteria does not take into account the area characteristics in terms of health, the modifiable areal unit problem (MAUP) may arise (Luo, 2013; Shi, 2009; Sloan et al., 2012) and the risk of aggregating areas with very different characteristics could emerge (Thompson et al., 2007).

Thus, the data must be sufficiently disaggregated to allow the researcher to perform the analysis, to obtain statistically robust results, and at the same time not compromising the individual confidentiality (Pearce et al., 2012).

Concerning the spatial risk factors of cancer described in the different papers considered in this review, it turned out to be difficult to identify which factors could promote cancer emergence, since the majority only provide some information and generally emphasise the need for further studies to confirm the results (Jemal et al., 2002). The reasons that could make it difficult to establish a relationship between cancer and spatial risk factors include: i) the latency period of the disease (Jarup et al., 2002; Toledano et al., 2001); ii) the situation in which a factor identified in one geographic region may not have the same effect in another region due to the presence or absence of other factors (Aragones et al., 2009); or iii) the fact that most cancers result from a combination of several factors rather a single one (Klassen and Platz, 2006).

Among the papers identifying cancer risk factors, we wish to highlight the following: i) the association between arsenic concentration in drinkable water and colon, lung and bladder cancer’s incidence risk increase in Cordoba, Argentina (Aballay et al., 2012); ii) a relation between higher incidence and mortality rates by cervical cancer, and more poverty and/or higher distance to screening in USA (Horner et al., 2011); iii) the urban disadvantage in risk of breast, colorectal, lung and prostate cancers in Illinois (McLafferty and Wang, 2009); or iv) a relation between Vitamin D insufficiency and an increase of prostate cancer risk (Schwartz and Hanchette, 2006); v) a possible association among coal mining activities and cancer mortality in West Virginia (Hendryx et al., 2010).

A lack of consensus was actually found in the papers found concerning cancer risk factors, and this was also extended to the methods applied. All articles presented here describe the methods and they vary considerably in many of them. A consensual issue is, however, the importance of creating maps that accurately describe disease spatial distribution patterns. This description can serve as a basis, firstly, for defining the areas in which more detailed studies on the disease aetiology must be carried out (Kullendorf et al., 2006) and, secondly, for identifying areas where interventions are needed to reduce the risk and mitigate the consequences of disease (Klassen and Platz, 2006).

Various methods, such as Bayesian models, Kriging, Spatial Scan Statistics and Moran’s I have been used in the analysis of spatial distribution patterns of cancer.

Bayesian approaches are sometimes criticised because of their failure to take into account the shape of the geographic areas under analysis. Some authors consider that these methods (Bayesian approaches) should be applied only when the shape of geographic areas are relatively homogeneous (Goovaerts, 2006a). If those areas are heterogeneous, it may be appropriate to use techniques that combine both global and local smoothing to deal with the inherent instability (Colonna, 2004). Other ways to eliminate the effect of this instability can be by applying tests of autocorrelation and spatial heterogeneity at the moment of choosing the Bayesian method (Colonna, 2004) or by using an adaptation of the Poisson Kriging, which includes analysing the size and shape of the geographical areas under study and the population density (Goovaerts, 2006a).

In linear generalised additive models (GAM) the predictor depends on unknown smoothing functions of some predictor variables, shifting the focus on inference about these making them a good alternative, in particular when the analysis includes the individual’s residential history (Vieira et al., 2008). The ability to incorporate the individual’s residential history is also a strong point noted in Q-statistics (Sloan et al., 2012).

With respect to cluster analysis, some authors consider it advisable to compare the results obtained by applying more than one method or software (Chen et al., 2008a), as this procedure allows a greater degree of certainty that one cluster corresponds to a real aggregation of cases (Bailony et al., 2011). Other authors go further and advise the comparison of results obtained by applying various methods at various scales in order to detect areas of activity in public health (Mohebbi et al., 2008). However, it should be remembered that the choice of the methods to be used depends in the end on the objectives of the study and the type of available...
data. More specifically, methods are not only closely related to the data under analysis, but also to their degree of geographic aggregation as well as to the features/factors influencing their distribution. For all of this, the most appropriate method for a given situation may not be the most convenient for another, although the analysis can be similar.

Finally, although a long search period was considered in this review, it is unlikely to include all published papers related to our subject. The choice of search criteria can have an impact on the final selection of papers, e.g., by requiring the presence of the word cancer in the title, which could lead to missing articles of possible relevance for the topic. Therefore, this paper does not exclude the need to be complemented with documents in official websites (e.g., IARC, 2016; NCI, 2016; NIH, 2016) regarding cancer data and research, as well as relevant papers and reports based on them (see, for instance, Ferlay et al., 2015 or Ryerson et al., 2016). Nevertheless, this review may help to promote research in this area, through the identification of some relevant knowledge gaps as well as the description and organisation of the knowledge based on the principal published literature. Moreover, cancer’s spatial epidemiology represents a very important concern, mainly for public health policies design aimed to minimise the impact of this chronic disease in specified populations.

Conclusions

Spatial epidemiology of cancer has been addressed in many articles, especially in the last decade, the most common cancer sites being breast, trachea-bronchus-lung, and prostate. Incidence rates were preferred over mortality rates as the epidemiologic frequency measure under study. Although individual data appear predominant in this review (74% of the articles), the units of analysis considered were geographic areas showing aggregation of cases. The research questions considered for analysis belonged to three different sets: i) spatial distribution and/or temporal evolution; ii) cancer risk factors; and iii) applied methods. The spatiotemporal evolution of cancer was covered in 50% of the papers analysis. The most common risk factors studied were demographic, socioeconomic and environmental. The methodological choice depended on data type and the analysis applied; specifically, the methodology was closely related to objectives, the data and their degree of geographic aggregation, including the features/factors influencing their distribution. This literature review comprised a large number of articles published in an extended period of years, which allowed presenting different approaches to spatial issues related to cancer epidemiology. Research on cancer’s spatial epidemiology represents a very important issue for decision-making and policies definition to fight one of the most important chronic diseases known.

References

by birth cohort. Spat Spatiotemporal Epidemiol 1:31-40.


Steinbock D, 2006. TagCrowd USA. Available from: tagcrowd.com


