Methodological challenges in cluster analyses dealing with rare diseases - A scoping review on childhood cancer studies.

Rossana Di Staso 1,2 Lorena Cascant Ortolano 3 Emilio Gianicolo 2

1 Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
2 Institute for Medical Biostatistics, Epidemiology and Informatics (IMBEI), University of Mainz, Germany
3 Departmental Library, University Medical Center Mainz, Johannes Gutenberg University Mainz, Germany

Objective

To improve our understanding of the prevailing techniques used for cluster-analyses of childhood cancer data, as well as the rationale behind choosing aggregate versus georeferenced data and the suitability of the control selection. This review forms part of a larger project analysing childhood leukaemia incident cases in Germany between 1990 and 2022.

Methods

A systematic literature review was conducted across MEDLINE (PubMed), Web of Science, and Scopus, from October 18, 2024, with no time limits. The search strategy combined MeSH terms and free-text keywords to ensure comprehensive coverage. Following retrieval, all records underwent rigorous deduplication using the SR-Accelerator Deduplicator tool. Studies applying geostatistical techniques to childhood cancer cases or mortality were included. After abstract screening, information was extracted from the eligible articles.

Results

A total of 2216 articles were screened, of which 72 were included (Figure 1). Of these, 31 used georeferenced data, 34 used aggregate data (mainly at the municipal level), and seven used a combination of both using different methodologies (Table 1). The majority of the studies had a focus on surveillance, while 23 investigated the role of a specific risk factor to explain the distribution of cases (e.g. incinerators).

Type of data	**n(%)**	**Type of cancer**	**n(%)**
Georeferenced data	31(43%)	Blood cancer	42 (58%)
Aggregate data	34(47%)	Solid cancer	8(11%)
Both	7(10%)	Both	22(31%)
Aim of the study		**Georef data methods**[1]	
Epidemiological surveillance	42(58%)	Knox test or K Function	26(68%)
Etiological hipothesis	23(32%)	Cuzick–Edwards test	6(16%)
Methodological intentions	7(10%)	Kernel density	2(5%)
Publication year		**Aggreg data methods**[2]	
≤1990	7(10%)	Indexes (Moran's, Tango's)	14(34%)
1990 - 2010	35(49%)	Poisson models	9(22%)
≥2010	30(42%)	Bayesian models	4(10%)
Patients		**Both data type methods**[3]	
Only children	64(89%)	Scan statistic	21(29%)
Several age classes	8(11%)		
Controls		**Outcome**	
Yes	10(14%)	ortality	3(4%)
No	62(86%)	Incidence	69(96%)

Table 1:Characteristics of the selected papers (N=72) as absolute number (N) and percentage (%). 1Percentage computed on the total number of paper that use georeferenced data (N=38); 2Percentage computed on the total number of paper that use aggregate data (N=41); 3Percentage computed on the total number of paper that use aggregate data (N=72);

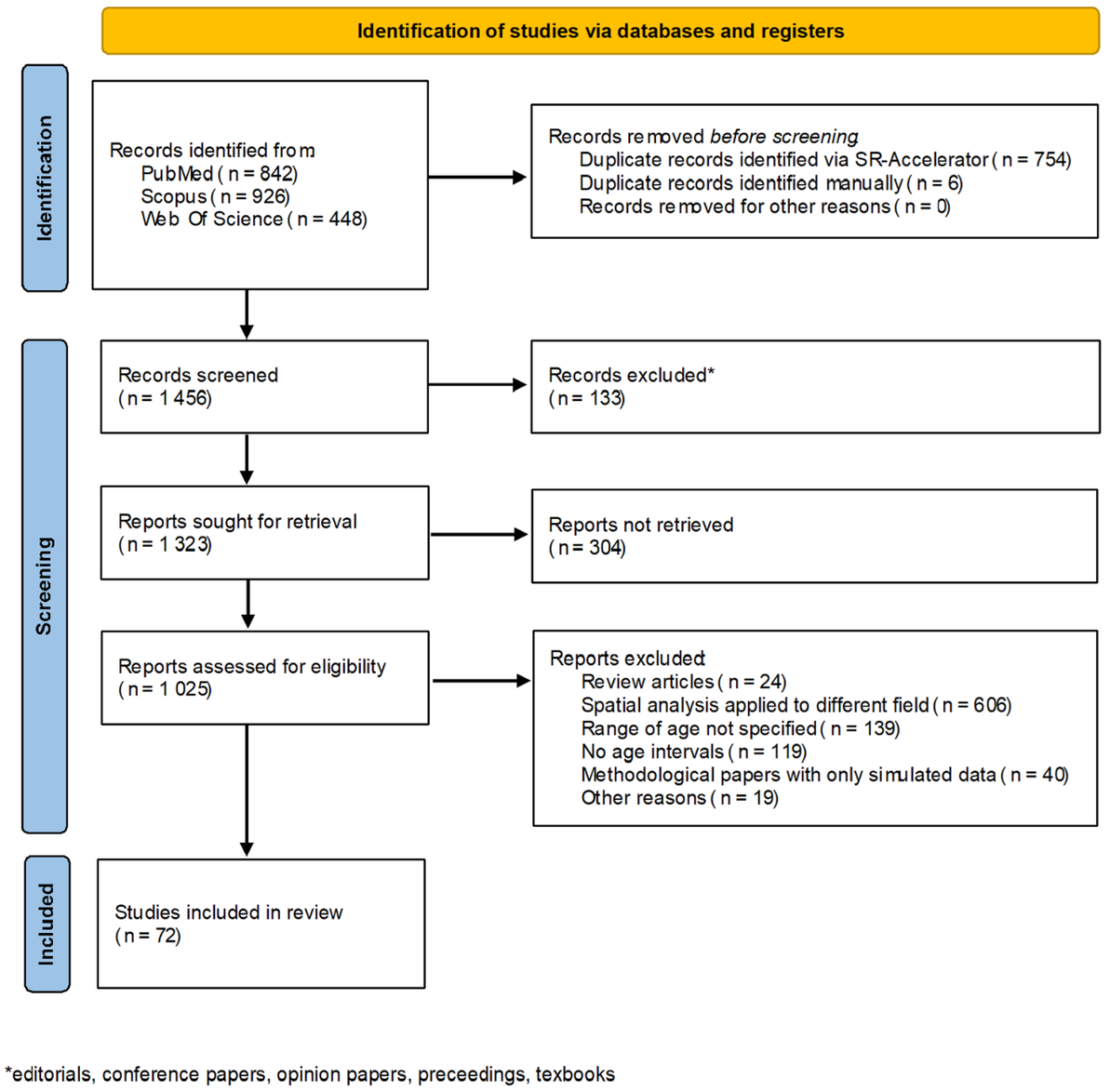

Figure 1:PRISMA flow diagram

Discussion

In most studies, only one method was applied and discussed. The Knox test for spatial/temporal clustering was the most common method for georeferenced data. This method can be used without a control group. However, all the authors highlight the arbitrariness of critical distances (defining the space and time clusters), which can be overcome to some extent by using the K-function. Only ten articles compared spatial distribution of cases with spatial distribution of controls. For aggregated data, the most commonly used method was the scan statistic with different window shapes and case distributions. This review was crucial in drafting an analysis protocol for leukaemia incidence in Germany.

Contact Information

- rossana.distaso2@unibo.it
- distasor@uni-mainz.de
- Rossana Di Staso - Via San Giacomo 12, Bologna (BO), 40126, Italy