

# APPLYING A SEQUENTIAL APPROACH IN ORDER TO DETECT AND INTERPRET A POSSIBLE INCREASED RATE OF MYELOID MALIGNANCIES IN THE GIRONA PROVINCE

**Enrique Y. Bitchatchi**

Per citar o enllaçar aquest document:  
Para citar o enlazar este documento:  
Use this url to cite or link to this publication:

<http://hdl.handle.net/10803/674092>

**ADVERTIMENT.** L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

**ADVERTENCIA.** El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

**WARNING.** Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.



**DOCTORAL THESIS**

**Applying a sequential approach in order to detect  
and interpret a possible increased rate of myeloid  
malignancies in the Girona province**

Enrique Y. Bitchatchi

2021



**DOCTORAL THESIS**

**Applying a sequential approach in order to detect  
and interpret a possible increased rate of myeloid  
malignancies in the Girona province**

Enrique Y.Bitchatchi

2021

**Supervisors**

Dr. Rina Chen, Dr. Elisabeth Pinart and Dr. Marc Yeste

**Tutor**

Elisabeth Pinart

Thesis submitted in fulfilment of the requirements for the degree of Doctor from the

University of Girona

Doctoral Programme in Molecular Biology Biomedicine and Health



“All scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge. That does not confer upon us a freedom to ignore the knowledge we already have, or to postpone the action that it appears to demand at a given time”

[Bradford Hill (1965)]

“Benefit of the doubt must go to the public not to the technology”

[Irwin D J Bross (1977)]

*A los afectados, que no se quedaron con nosotros  
porque no prevenimos a tiempo*

*To those affected, who didn't stay with us  
because we were unable to prevent it in time*

## ACKNOWLEDGEMENTS

Ms. Antonia Mora, together with the library keeper and co-workers for taking devoted care of the cleanliness and tidiness of the facilities in the library of our campus Montilivi. The attentive staff never sparing an amiable service in the cafeterias of the Science and Law Faculties on the campuses of Montilivi and Emili Grahit. That has been very kind of you.

Thank you to Mr. Joan-Carles Corney (Emili Grahit UdG library) for his kind assistance as documenter, notwithstanding the misleading algorithms and subject headings the National Library of Medicine accounts for the disease clustering realm. I am so grateful to the library personnel and especially to Mr. Ramon Sánchez, the unassailable librarian for all his assistance, anytime. Also, I want to thank Ms. Roser Benavides for her guidance throughout the reference manager.

Some 40 years ago Dr. José Esparza supervised my assistantships at the Centre of Microbiology and Cell Biology from the Venezuelan Institute for Scientific Research. Let me attest that José imparted a contagious irremediable passion for research in public health just as an inclination towards competing explanations for evidences – featuring an R<sup>o</sup> of tens. Those days I was a newcomer in clinical medicine; I have José to thank for initiating me into the dwarf-lettered volumes of *Excerpta Medica*. It was just for me to challenge a medical teacher of mine who wallowed in claiming an (unproven) causal association between gastric ulcer disease and bad mood. No sooner had the entitled physician crippled with dyspepsia engendered by a photocopied paper from *The Lancet* than he abandoned his arrogance.

I must thank Mr. David Matas and Mr. Albert Puig, the informatics support at the library. To Dr. Joan Solé and Dr. Germà Coenders for that timely invaluable solidarity and guidance as well as to that help earnestly encountered with professors Joan Saldaña, Natalia Adell and Marc Corfu, at the points where I got stuck. My sincere thank to the internal and external appointed reviewers of my monograph just as the ones and the Associate Editor who provided with their enlightening remarks that much improved our article published in CANEP. I do appreciate the support of the OITT and GRECS personnel proactively responsive to any question in their remit.

I wish to thank the graphic designer Ms. Francisca A. Sotomayor for her drawings and the artistic polishing of my slides. I am grateful to Mrs. Maria J. Kistic who accompanied me through the thorny way of getting ready for the defence of my dissertation.

Thank you to my supervisors and tutor Marc and Beth for your generous support throughout the dissertation process, and to my senior supervisor, the late Rina towards whom I am committed as a follower of her pursuits. You let me break the odd bone and I owe you so much, *Maestra*.

My last, but never least, Thank You goes to my lover for being there and to my family for everything.

Finally, to a large extent, my project, my sociomoral choices to this day come from that magma to which I exposed myself as a student at UCV – Alma Mater, Venezuela.

## **DERIVED PUBLICATION**

Rina Chen & Enrique Y. Bitchatchi.

Detection and estimation of the increasing trend of cancer incidence in relative small populations.

*Cancer Epidemiol.* 50, 207–213 (2017).

DOI: 10.1016/j.canep.2017.04.005

Impact Factor 2.179, Q2 (Epidemiology, Cancer Research, Oncology; 2019

<http://www.scimagojr.com/index.php>.)



## **LIST OF ABBREVIATIONS**

ALL	acute lymphocytic leukaemia
AML(s)	acute myeloid leukaemia(s)
ATSDR	Agency for Toxic Substances and Disease Registry
CML	chronic myelogenous leukaemia
CMML	chronic myelomonocytic leukaemia
CUSCORE	cumulative score technique
CUSUM	Cumulative sum technique
ET	essential thrombocythaemia
GCR	Girona Cancer Registry
ICD-O-3	International classification of diseases for oncology (3 <sup>rd</sup> edition)
IARC	International Agency for Research on Cancer
JAK2	Janus Kinase 2 gene mutation
LH(s)	lymphohaematopoietic or haematolymphoid malignancie(s)
MDS	myelodysplastic syndromes
MDS–MPN	myelodysplastic/myeloproliferative neoplasms
MM(s)	myeloid malignancie(s) as per <i>ICD-O-3</i> morphology codes
MPN	myeloproliferative neoplasms
O/E	observed:expected (ratio)
PMF	primary myelofibrosis
PV	polycythaemia vera
Qi	q-interval
RI	relative interval
SIR	standardized Incidence Ratio
SMR	standardized Mortality Ratio
SQ	accumulative q interval

## **CONTENTS**

ACKNOWLEDGEMENTS .....	V
DERIVED PUBLICATION .....	VII
LIST OF ABBREVIATIONS .....	VIII
LIST OF FIGURES .....	XII
LIST OF TABLES .....	XIII
ABSTRACT .....	1
THESIS BLUEPRINT .....	5
1. INTRODUCTION.....	6
1.1. Why bother?.....	6
1.2. Cluster and clustering definition .....	10
1.3. Nature of temporal chronic disease clusters.....	12
1.3.1. The low frequency attribute .....	12
1.3.2. Actual or spurious incidence variation.....	13
1.3.3. Appropriateness of event data and denominators.....	14
1.3.4. False positive and false negative errors and clusters.....	15
1.3.5. The One-off feature.....	17
1.3.6. Clusters embedding in the series .....	18
1.4. Techniques for temporal cluster inquiries .....	20
1.4.1. Sequential techniques for temporal cluster inquiries.....	21
1.4.2. Playing characters .....	24
1.4.3. The alarm-signalling test. The cuscore's basics .....	28
1.4.4. Confirmatory procedure .....	28
1.4.5. Graphical display by accumulative q-interval (sq).....	28
2. THE STUDY. THE ANCILLARY CONNECTIONS .....	30
2.1 Clusters of lymphohaematopoietic malignancies in communities .....	31
2.2 Myeloid Malignancies.....	32
2.2.1 Commonalities .....	33
2.2.2 Constitutional proneness and acquired susceptibility.....	34
2.2.3 Exogenous causes.....	35
2.3 Nomenclature and leukaemogenic clustering.....	38
3. RESEARCH QUESTIONS AND OBJECTIVES .....	41
3.1 Research questions .....	41
3.2 Objectives.....	43
4. MATERIALS AND METHODS .....	44
4.1 Outline.....	44

4.2	Reference Population .....	45
4.3	Myeloid Malignancy cases.....	45
4.3.1	Descriptive analyses on person data.....	46
4.3.2	Subtypes' prevalences .....	47
4.4	Expected morbidity .....	47
4.4.1	Test of significance for SIRs .....	48
4.5	Observed interval as random variable and RI appraisals .....	48
4.5.1	Tailoring RIs .....	49
4.5.2	The alarm signalling technique .....	50
4.5.3	Handling incomplete recording of dates .....	51
4.6	Censored intervals .....	52
4.7	Confirmatory procedures.....	53
4.8	Graphical display of temporal pattern .....	54
4.9	Software .....	55
4.10	Quality control .....	56
4.11	Bioethics.....	56
5.	RESULTS.....	57
5.1	La Selva County.....	58
5.1.1	Santa Coloma de Farners.....	58
5.1.2	Sant Hilari Sacalm.....	62
5.1.3	Arbúcies .....	64
5.1.4	Blanes.....	67
5.1.5	Lloret de Mar.....	67
5.2	Gironès County .....	68
5.2.1	Girona town.....	68
5.2.2	Cassà de la Selva .....	73
5.2.3	Salt.....	75
5.3	Pla d'Estany County.....	76
5.3.1	Banyoles .....	76
5.4	Baix Empordà County.....	81
5.4.1	Palafrugell .....	81
5.5	Ripollès County.....	83
5.5.1	Ripoll.....	83
5.6	Concurrent independent clusters .....	84
5.7	Predominant subtypes in clustered communities .....	85
6.	DISCUSSION .....	89

7. CONCLUSIONS .....	109
8. REFERENCES .....	110
9. APPENDIX .....	123
APPENDIX 1 .....	123
APPENDIX 2 .....	130
APPENDIX 3 .....	131

## **LIST OF FIGURES**

<b>Figure 1. Annual amount of IARC Group 1 carcinogen emissions in the proximity of 8098 Spanish towns (2007-2010).....</b>	<b>10</b>
<b>Figure 2. Observed and expected cumulative q-intervals for AML in Santa Coloma de Farners (1994-2008).....</b>	<b>60</b>
<b>Figure 3. Observed and expected cumulative q-intervals for MDS in Santa Coloma de Farners (1994-2008).....</b>	<b>61</b>
<b>Figure 4. Observed and expected cumulative q-intervals for MPN in Santa Coloma de Farners (1994-2008). ....</b>	<b>63</b>
<b>Figure 5 Observed and expected cumulative q-intervals for MPN in Sant Hilari Sacalm (1994-2008).....</b>	<b>66</b>
<b>Figure 6 Observed and expected cumulative q-intervals for MPN in Arbúcies (1994-2008).....</b>	<b>67</b>
<b>Figure 7 Observed and expected cumulative q-intervals for MDS in Girona town (1994-2008).....</b>	<b>70</b>
<b>Figure 8 Observed and expected cumulative q-intervals for MPN in Girona town (1994-2008).....</b>	<b>73</b>
<b>Figure 9 Observed and expected cumulative q-intervals for MDS in Cassà de la Selva (1994-2008).....</b>	<b>75</b>
<b>Figure 10 Observed and expected cumulative q-intervals for MPN in Cassà de la Selva (1994-2008).....</b>	<b>76</b>
<b>Figure 11 Observed and expected cumulative q-intervals for AML in Banyoles (1994-2008).....</b>	<b>78</b>
<b>Figure 12 Observed and expected cumulative q-intervals for MDS in Banyoles (1994-2008).....</b>	<b>80</b>
<b>Figure 13 Observed and expected cumulative q-intervals for acute myeloid leukaemias in Palafrugell (1994-2008).....</b>	<b>83</b>

## **LIST OF TABLES**

<b>Table 1 Cases per community-year and Observed:Expected ratios for 15 preselected municipalities 1994-2008.....</b>	<b>59</b>
<b>Table 2 Expected and observed number of AML, MDS and MPN diagnoses, Santa Coloma de Farners.....</b>	<b>60</b>
<b>Table 3 Sequential assessment, CUSCORE. MPN subjects, Santa Coloma de farners.....</b>	<b>62</b>
<b>Table 4 Sequential assessment. MPN subjects, Sant Hilari de Sacalm.....</b>	<b>65</b>
<b>Table 5 Sequential assessment, CUSCORE. MDS subjects, Girona town.....</b>	<b>70</b>
<b>Table 6 Sequential assessment, CUSCORE and confirmation. MPN subjects, Girona town.....</b>	<b>72</b>
<b>Table 7 Sequential assessment, AML subjects, Banyoles.....</b>	<b>77</b>
<b>Table 8 Sequential assessment, MDS subjects, Banyoles.....</b>	<b>79</b>
<b>Table 9 Sequential assessment, AML subjects, Palafrugell.....</b>	<b>82</b>
<b>Table 10 By-gender O /E by main MM category at clustered communities 1994-2008.....</b>	<b>83</b>
<b>Table 11 Period prevalence by malignancy subtypes, communities with and without clustering 94-08...88</b>	<b>88</b>





## ABSTRACT

**Background:** Based on preventive and precautionary dictates, the availability of procedures to detect increased morbidity rates in relatively small populations or clusters of rare diseases stands paramount. In this context, cost-aware, instructive and timely analyses must provide guidance thereby determining whether cases relate causally and the nature of their causal mechanisms. In addition, these analyses must indicate whether further assessments are needed and/or if any active environmental intervention in the pertinent community is required. Genuine aggregates of infrequent chronic diseases may be rendered biased or unnoticed by just measuring general secular trend statistics. Discretionary surveillance of these conditions entails the appearance of incidental clusters and costly supervening elucidation. In the foreground of the present study lies a report of ascending secular trends for leukaemias of myeloid lineage (MMs), in a province of 730,000 people. This warrants an inquiry plan aimed at unveiling causative clues. The main aim of this thesis is to demonstrate the use of a sequential-based approach in detecting an elevated rate that may have initiated at some unknown point of time during the study period.

**Objectives:** 1) To assess the temporal clustering of MMs through sequential procedures by waiting time. 2) To probe into the intensity, time pattern and overall place-dependent behaviour of any detected cluster. 3) To inquire into person variables to help informing possible causative paths in the communities bearing indicative clustering.

**Methods:** Analyses were based on the registered cases of the main MM categories occurring at preselected municipalities of Girona Province (Spain) over a 15-year period. Every municipality with at least 10 diagnoses for some category was selected; this totalled 35 series in 15 communities. Subtle clustering was validated using a cumulative score test (CUSCORE) that signals an alarm for any stretch during the analyzed diagnoses. A CUSCORE test hinges on the *Relative Interval (RI)* statistic, which reflects the waiting-time-to-event; this constitutes a continuous variable that controls for size and profile differences between populations and sub-periods. Following the CUSCORE test, a graphical display of the temporal pattern and a confirmation test were conducted. These procedures were preceded by ascertaining standardized incidence ratios (*SIR*). Occurrences of MM-subtypes within the temporal agglomerations of cases were assessed too.

**Results:** Eighteen series (11 communities) evinced excess of observed MM cases by the 15-year *SIR*. The *RI*-based sequential procedures unveiled the temporal patterns of the clusters over the multiyear period. Their detection yield even proved at sub-unity event counts per year and at mild-intensity epidemic rises. Six communities registered one or more indicative time clusters. There is no reasonable chance of observing more than one cluster of those in each community during 15 years. Sometimes 2 MM-categories overlapped for the epidemic spell. Depictions of waiting intervals evidenced embedded huddles within the series. Once a stretch of these 10 clusters was sensed, the focused analyses on morphological subtypes showed selective involvement by *de novo* subtypes in 8 of them. Some causative hints emanated from observed deviations from usual gender ratios or ages at diagnosis in the clustering communities. And so did 'within-cluster' subtype frequencies that were divorced from anticipated prevalence.



**Conclusions:** Addressed beforehand, upon paucity of occurrences, these sequential procedures worked usefully for the ad hoc test of alarm signal, post alarm deciphering of time pattern, and cluster confirmation. Where the descriptive epidemiology failed to show an increased secular trend, the aforementioned methods uncovered small community-based clustering and helped in tailoring realistic intensity estimates related to unobtrusive ballooning rates. Whereas the integrated assessments used herein focused on cancer agglomeration, this rather economical approach could be extended to other non-return maladies if high quality morbidity data derived from population-based registries are available. The insights, achieved as plausible causation understandings of the signalled clusters, hint what ought to be investigated at the next elucidation phase, and outreach to mending interventions and pre-empting preventable chronic maladies.

---

**Antecedentes:** Ineludiblemente, principios de prevención y de cautela requieren acceso a procedimientos capaces de detectar clústeres genuinos de enfermedades infrecuentes o tasas incrementadas de morbilidad en pequeños grupos de población. Deben garantizar el análisis instructivo, expeditivo, y coste-consciente. Habrán de orientar sobre la existencia de relación causal entre los casos agregados, la previsible naturaleza del mecanismo causal evidenciado por los datos, la perentoriedad de investigación ulterior o bien de intervención medioambiental en la comunidad pertinente. La mera medición de tendencias seculares de tasas correspondientes a enfermedades infrecuentes de no-retorno, puede falsear o bien fracasar en detectar una agregación genuina de éstas. La monitorización discrecional de este tipo de condiciones, implica la emergencia de clústeres incidentales y un subsiguiente derroche de recursos dedicados a su dilucidación. En el primer plano de esta tesis subyace un reporte de tendencias seculares incrementales de leucemias mieloides (MMs) en una provincia de 730.000 habitantes. Esto demanda una estrategia de investigación para develar claves causales. El propósito principal de esta disertación es demostrar la adecuación de un abordaje secuencial para detectar una tasa elevada que pudiere haberse iniciado en algún momento desconocido durante el período de estudio.

**Objetivos:** 1) Evaluar clusterización temporal de MMs mediante procedimientos secuenciales por tiempo de espera. 2) Examinar la intensidad, el patrón de tiempo y el comportamiento general dependiente del lugar de cualquier grupo detectado. 3) Explorar variables de género, edad y subtipo de leucemia para ayudar a informar las posibles rutas causales en las comunidades con clústeres indicativos.

**Métodos:** Se analizaron comunidades preseleccionadas de la Provincia de Girona (España), para las categorías principales de MM durante 15 años. Se escogieron todos los municipios que registraron  $\geq 10$  diagnósticos para alguna categoría, totalizando así 35 series dispersas en 15 comunidades. Las clusterizaciones inadvertidas se validaron mediante la prueba CUSCORE: una puntuación acumulativa que señala una alarma en cualquier segmento durante los diagnósticos analizados. CUSCORE depende del estadístico *Intervalo Relativo (RI)*, el cual refleja el tiempo de espera hasta un evento. A su vez constituyendo una variable continua que corrige las diferencias en tamaño y estructura entre poblaciones y entre subperíodos. Subsiguiendo al test CUSCORE, se ejecutaron trazados gráficos de patrones temporales y un test de confirmación. Estos procedimientos fueron precedidos por valoraciones de razones de incidencia



estandarizadas (*SIR*). Dentro de las agregaciones temporales, las frecuencias de subtipos de MMs y variables demográficas a nivel de individuo también fueron evaluadas.

**Resultados:** Dieciocho de las series (11 comunidades) de 15 años, mostraron un exceso de MMs observados (*SIR*). Los procedimientos secuenciales basados en *RI* expusieron los correspondientes patrones temporales de los clusters durante el período multianual y mostraron un rendimiento para su detección, aún ante recuentos de eventos  $< 1$  anual y expansión moderada de intensidad epidémica. Seis comunidades soportaron uno o más clusters. Es razonablemente improbable hallar más de un clúster por comunidad en 15 años. Ocasionalmente concurren dos categorías de MM en el lapso epidémico. Representaciones de intervalos de espera evidenciaron 10 aglomerados inmersos dentro de las series, que una vez definidos sus confines, concentraron subtipos morfológicos *de novo* predominantes en 8 de aquellos. Algunos indicios causales emanaron de las desviaciones observadas de las proporciones de género habituales o de las edades en el momento del diagnóstico en las comunidades con clústeres, así como de variabilidad en las frecuencias ‘intraclúster’ entre los subtipos de estas leucemias.

**Conclusiones:** Estos procedimientos secuenciales funcionaron como sensores fiables *ad hoc* de alarma, aportando un descifre de los patrones temporales y prueba post-alarma de confirmación de los clusters. Ello, ante incidencias bajas en series acometidas *a-priori*. Mientras que con epidemiología descriptiva se descartaba erróneamente la hipótesis de tendencia secular incremental, nuestros métodos señalaban clusterización a nivel de pequeña comunidad y asistían en la revisita de la intensidad de ascensos inadvertidos de tasas. Si bien tales procedimientos económicos sirvieron aquí para evaluar aglomerados de cáncer, su utilidad es generalizable a otras enfermedades de no-retorno; aunque condicionado a la asequibilidad de datos derivados de registros de incidencias de base poblacional. La plausibilidad de las derivaciones de causalidad a partir de clústeres verosímiles guía a lo que ha de ser investigado en la fase dilucidadora ulterior, con alcance a intervenciones reparadoras y evitación de dolencias crónicas prevenibles.

**Bagatge:** Ineludiblement, els principis de prevenció i de cautela requereixen de la disponibilitat de procediments capaços de detectar clústers genuïns de malalties infreqüents o taxes incrementades de morbiditat en petits grups de població. Aquestes han de ser anàlisis instructives, expeditives i cost-conscients, i han d'orientar sobre l'existència de relació causal entre els casos agregats, la previsible naturalesa del mecanisme causal evidenciat per les dades, la preemtorietat de recerca ulterior o bé d'intervenció mediambiental en la comunitat pertinent. La simple determinació de tendències seculares de taxes corresponents a malalties rares de no-retorn pot falsejar o bé fracassar en la detecció d'una agregació genuïna d'aquestes. El monitoratge discrecional d'aquest tipus de morbiditat, implica l'emergència de clústers incidentals amb el conseqüent malbaratament de recursos. Al capdavant d'aquesta dissertació jeu una publicació que presenta tendències seculares incrementades de leucèmies mieloides (MMs) en una província amb 730000 residents. Això exigeix una estratègia de recerca per a desxifrar claus causals El propòsit d'aquesta tesi doctoral és demostrar l'adequació de seqüències temporals eficients davant recomptes baixos de casos escampats durant períodes prolongats.

**Objectius:** 1) Avaluar clusteritzacions temporals de MMs mitjançant procediments seqüencials per temps d'espera. 2) Examinar la intensitat, el patró de temps i el comportament general dependent del lloc de qualsevol grup detectat. 3) Explorar les variables de gènere, edat i

subtipus de leucèmia per a ajudar a informar les possibles rutes causals en les comunitats amb clústers indicatius.

**Mètodes:** Es van analitzar comunitats preseleccionades de la Província de Girona (Espanya), per a les categories principals de MM durant 15 anys. Es van triar tots els municipis que van registrar  $\geq 10$  diagnòstics per a alguna categoria, totalitzant així 35 sèries disperses en 15 comunitats. Les clusteritzacions desapercebudes, es validaven segons el test CUSCORE una puntuació acumulativa que assenyalava una alarma en qualsevol segment durant els diagnòstics analitzats. El test CUSCORE depèn de l'estadístic *Interval Relatiu (RI)*, el qual reflecteix el temps d'espera fins a un esdeveniment; aquest constitueix una variable contínua que corregeix per diferències en grandària i estructura entre poblacions i entre subperíodes. Subseguint el test CUSCORE es van executar traçats gràfics de patrons temporals i un test de confirmació. Aquests procediments van ser precedits per valoracions de raons d'incidència estandarditzades (*SIR*). Dins dels agregats temporals, les freqüències de subtipus de MMs també van ser avaluades.

**Resultats:** Divuit de les sèries (11 comunitats) de 15 anys, van mostrar un excés de MMs observats (*SIR*). Els procediments seqüencials basats en RI exposaren els corresponents patrons temporals dels clústers durant el període pluriennal. Indicaren un rendiment per a la seva detecció, fins i tot en el recompte d'esdeveniments  $< 1$  anual i expansió moderada d'intensitat epidèmica. Sis comunitats van tenir un o més clústers, i ocasionalment van donar-se dues categories de MM en el lapse epidèmic. Ací, és raonablement improbable trobar més d'un clúster per comunitat en 15 anys. Les representacions d'interval d'espera van evidenciar 10 aglomerats immersos dins de les sèries, que un cop definits els seus límits, concentraren subtipus morfològics *de novo* predominants a 8 de aquells. Les desviacions de les raons per sexe o de l'edat mediana de diagnòstic a les comunitats afectades, així com alteracions en les freqüències 'intraclúster' entre subtipus, proveeixen certes claus de causalitat.

**Conclusions:** Aquests procediments seqüencials funcionaren com a útils sensors fiables ad hoc d'alarma, i van permetre desxifrar els patrons temporals i prova post-alarma de confirmació dels clústers. Això, davant d'incidències baixes en sèries escomeses a-priori. Mentre que amb l'epidemiologia descriptiva es descartava erròniament la hipòtesi de tendència secular incremental, els nostres mètodes permeten identificar la clusterització a nivell de comunitat petita i permeten revisar la intensitat d'ascensos inadvertits de taxes. Malgrat que aquests procediments econòmics van servir per a avaluar aglomerats de càncer, la seva utilitat és generalitzable a altres malalties de no-retorn; per bé que condicionada a l'assequibilitat de dades d'alta qualitat derivats de registres d'incidències de base poblacional. La plausibilitat de les derivacions de causalitat a partir de clústers versemblants guia al que ha de ser investigat en fase d'elucidació ulterior, amb abast d'intervencions reparadores i prevenció de malalties cròniques prevenibles.



## THESIS BLUEPRINT

The present thesis dissertation is divided into six sections and a final part for concluding remarks. It also provides an abstract and its translation into Catalan and Spanish. A subsidiary article to the core thesis and published in *Cancer Epidemiology* is appended.

The **first section** is an introduction expounding on the scope of temporal clustering of chronic no-return maladies. Here I address subtle epidemic intensities at the same time in parallel with of the realm of low frequencies/rare sickness. This opening frames attendant methodological issues and hindering factors of inference – typically leading us to plod through thorny endeavours. You hear the hoofbeats: zebras (causal clusters) or horses with stripes (e.g. rather chance aggregations)? Much as the ‘why bother’ hesitation, so long as it behoves health authorities to minimize delay -not least to ensure accountability and transparency for all stakeholders- defaulting to idleness will not stand up to the scrutiny of the finalistic tenet immanent in environmental epidemiology.

In the lead up to the case for action, a due underlying gamma, memoryless distribution of the concerned morbidity data marks the point to undertake sequential methods. I then introduce the sequential techniques i.e. the character’s attributes, which I am going to launch en bloc in my research.

In the **second section**, I put forward the pre-hoc study to be tackled. Namely, the secular trends upturn in myeloid lineage leukaemia timely reported by researchers from the population-based cancer registry in the province of Girona. As in most real instances where a somewhat unfocussed concern about a population’s health arises, the retrospective data set does not include a control group. The tendencies showed a subtle epidemic. Besides, we were primed with no background to the carcinogenic environments. Hence, one cannot know beforehand when the risks started to rise. Therefore the type of analyses for you to marshal here turned out to be of sequential type by the time-to-wait.

The cancer registry morbidity data suited high quality, fitness-for-use standard contributed by an utmost coverage across the province catchment area in its entirety, as well as by an exhaustive coding according to WHO classification consensus. In addition, the registry had performed a taxonomical harmonization throughout. Therefore it was assumed worthwhile to engage myself in seeking for best settled knowledge that might inform causative elucidation about some revealing rate increases; by revolving around main categories as well as fine-tuned subtypes among these neoplasias.

The **third section** develops the research questions for this non-ex post investigation. Should there be any perceived clusters, I anticipate the spectrum of outputs that might be pinpointed with the launching of the implied integrative approach. This consists of: an overall count or a sequential test, the sensing of the time pattern and risk-fold; a confirmatory procedure to reduce false positive rates, and a graphical depiction that pivots on a standardized statistic – which enables us to unfold huddles that otherwise will remain unearthed. And secondly -so as not to waste useful nuggets of information- I expound upon causative elucidation that can be produced with inferences drawn from the aforementioned pathophysiological background as

well as from age and gender variables: most namely their departure from the anticipated in affected communities.

This binds us to formulate 3 objectives: 1) To assess the temporal clustering of MMs through sequential procedures by waiting time. 2) To probe into the intensity, time pattern and overall place-dependent behaviour of any detected cluster. 3) To inquire into person variables to help informing possible causative paths in the communities bearing indicative clustering.

The **fourth and fifth** sections entail the main contributions. I analyze the complete sample of the municipalities of the province where at least 10 individuals have suffered from one of the 4 categories of myeloid malignancies. This amounts to about half the diagnosed persons during 1994-2008. Then an a priori-settled analysis is executed on each series and by each community. The identification of concurrent clustering, sound person variable deviances, subtype predominance in clustered communities and its risk-folds, as well as any noticeable geographical neighbourhood between communities with temporal clustering helps us to ponder in favour -or not- of the case for the genuineness of an indicative cluster and attendant causative and temporal hints.

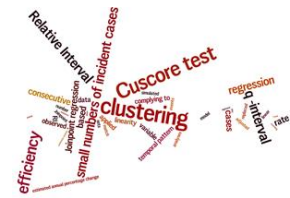
The last couple of sections (**6<sup>th</sup> and 7<sup>th</sup>**) have been devoted to discussing and to summing up the appropriateness, downsides and strengths of the integrative approach present. I expand on its compliance with the cornerstones that benchmark agencies deem adequate for clustering research tools in chronic illnesses. Here its far-reaching utility is proposed.

The appended article entitled Detection and estimation of the increasing trend of cancer incidence in relatively small populations, is a complementary study. This has been inspired by a contemporary report from the same cancer registry about the rise of an uncommon distinctive myeloid malignancy. We explicitly appreciate the need to address the (weak) efficiency of the regression procedure thereof: How powerfully it leads to the detection of an incidence trend of infrequent diseases? Would a waiting-time test outperform? The paper has been aptly published in a special issue of *Cancer Epidemiology* devoted to addressing cancer in small states.

## 1. INTRODUCTION

### 1.1. Why bother?

From an ethical point of view, environmental epidemiology, and medicine in general, are not neutral scientific disciplines, but finalistic activities. The main purposes of these two disciplines are to not only “study the nature” but also to promote well-being, prevent harm and minimize damage to ecosystems.



Health authorities have to address queries of increased incidence of infrequent or rare illnesses. The appeal *stares you in the face*; or, as the editorial team in *The Lancet* (1990), pronounces it <<the epidemiologist's nightmare>><sup>1,2</sup>: in fact, health officials are exhorted to address the appeal and make defensible decisions. Yet, this positioning expects for setting detection procedures aimed at addressing whether or not should one proceed to mending interventions. And if not, to pilot through more expensive and exhaustive research undertakings. Likewise, responsible authorities will undertake a throughout and informative epidemiologic or laboratory research for the purpose of learning about specific factors in the occurrence of the maladies in question <sup>3</sup>.

Decision makers in public health policy though aim at solving problems related to outcomes ranking higher on socioeconomic impact: a political paradigm where infrequent ill conditions seldom qualify <sup>4-6</sup>. However, setting public health priorities up rather obliges a revisited perspective.

To make the point let us expound on the truism-like statement, according to which, improvements made on the understanding of no-return-illnesses occurrence or causation may well provide transcendental policy clues <sup>7</sup>, as follows: First, by entering into neglected research realm, one can find new potential risk determinants.

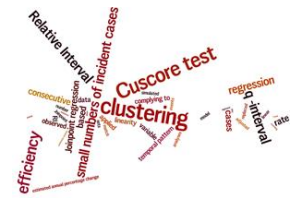
Second, upon investigating uncommon diseases, risk factors of importance for other maladies may be revealed, never precluding analogue hazards. Think of malignancies: the realm of cluster detection and investigation usually expands beyond the malady eventually signalled <sup>8</sup>. Carcinogens often evince health disruptions other than tumoral. For example, the Woburn disaster stemmed from 20 cases of childhood haematolymphoid malignancies (**LHs**) over 20 years. The Massachusetts Department of Public Health moved beyond these malignancies to the analysis of the incidence of adverse pregnancy outcomes plus 14 categories of childhood disorders. However, no one had suggested that congenital abnormalities or any reproductive conditions would be related, as yet, to the Woburn contaminated wells. A tenable correlation

between carcinogenicity, teratogenicity and other adverse reproductive effects endorsed a case for action of pursuing so. This approach has never been challenged even though, only the cancer cluster attained final association to the tainted water <sup>9</sup>.

Returning to malignancies, a consequence of sensitivity-enhanced research is that *additional cancer sites* become ascribed to established carcinogenic determinants. This is the case when realizing that a high toll-taker cancer site can be attributable to a causative agent that had been only reckoned with less ubiquitous malignancy <sup>10</sup>. Thus, gains in public health priorities might in consequence evolve.

Third, cluster investigations do warrant far-reaching interventions, and even exploratory modalities of cluster analyses entail public health actions or interventions <sup>5,11</sup>. At the end of the day, interacting with a concerned community opens up a wide opportunity for health education on preventable diseases <sup>2,8</sup>; it may serve as a basis to appeal for either reinforcing detailed exposure monitoring (and clean-up if need be – if the mess still lasts), or assembling the at-risk-population files <sup>12</sup>. An example for the vast scope of monitoring, as assumed by the Spanish Government, appeared on the media in 2017. The new spread a study performed by the Institute of Health Carlos III that leans on the Register of Emissions and Pollutant Sources. Aptly, a Continental command issued to frame a *confident use* of potentially harmful agents with likely long-term chronic effects (Figure 1) <sup>3,7,13,14</sup>.

Failure to comply with such a commandment alienates the public by exhausting their trust towards the governmental and the industries' corporates, and the procedures employed to prevent disasters and hazards. Paraphrasing Richter & Laster (2004) <sup>15</sup>, if disappointing the enforcement of legal tools such as *right to Know*, *right to act*, *informed consent*, the misuse of universal ethical norms supervenes. Nobel Memorial Prize in Economic Sciences J.E. Stiglitz spotlighted similar neglect at the national-governmental context. By dismissing the social contract, he stated, citizens may stop complying with the reciprocal contracts and their ones with the Administration <sup>16</sup>. Unlawfulness leaks into the public framework and wreaks havoc on



the core of social interrelationships, as puts a scholar in governance and citizen security <sup>17</sup>. Definitely, we concern ourselves with because *we all* entail such reliability.

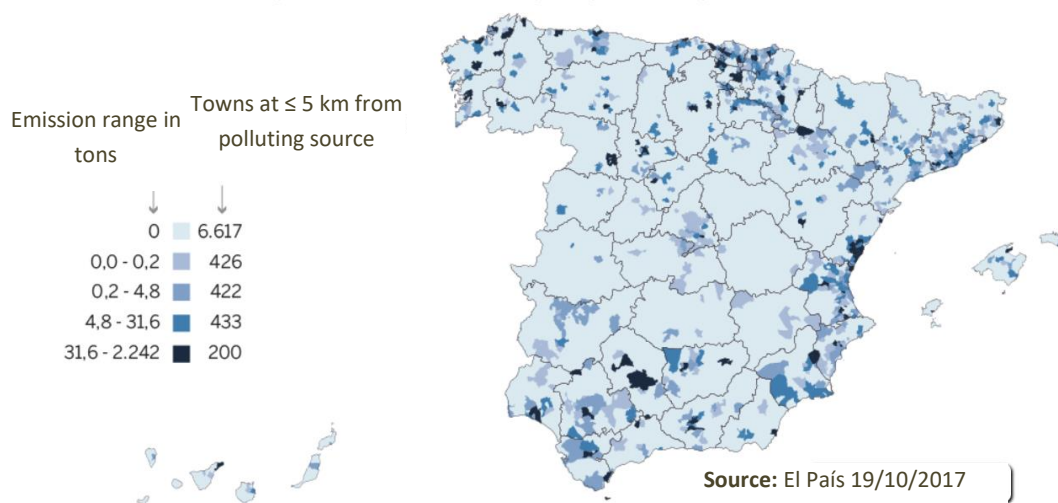
Fourth, for these interventions embed another outflow. That is, the recruit of knowledgeable (outsider to the responsible agencies) as well as alert laymen <sup>12,18</sup>. The revisited CDC guidelines (2013) set to investigate clusters put the involvement of communities forward <sup>11</sup>. The tainted-water-disaster at Marine Corps Base Camp Lejeune, North Carolina exemplifies how public health researchers comply with the ethos of taking actions on community participation <sup>19</sup>. So much so, these scientific experts recommend the creation and then manage to firmly exchange and work with a Community Assistance Panel, made up of members of the base community <sup>19,20</sup>.

Finally, *no action* is not an option <sup>21</sup>. An absence of clear proof of damage should not be allowed to lead to passiveness towards detection of clusters or to any relief of effort to secure higher standards of control. A great deal of what is currently known of causation in occupational malignancies stems from *uncommon* signal neoplasms, which have been identified only because they *clustered* in a particular set of workers. Health workers as well as other stakeholders are entitled to take pains over raising the question *Are there other affected ones?*, viz., the highest cost-effective interview tool in social medicine. Since the 18<sup>th</sup> century, several diseases that ensued to be connected to jobs and workplaces -and aggregated as genuine clusters- had been originally called attention to by incisive workers or honest physicians <sup>22</sup>.

The <<*Why bother*>> question recollect passiveness. Passiveness would not lead the discovery of AIDS, the most dramatic epidemic of contemporary times. AIDS story evolved from a prime cluster report of five cases in Los Angeles County. There, the plain scrutiny of routine morbidity reports channelled the revelation of an uncommon aggregation of events. Notwithstanding, we might as well remember that just charting and timely reporting of a seeming excess of events is insufficient to avoid the precipitation of morbidity death casualties or harm to wellbeing; not to mention the prospect for some relief from a deepening inequality in



the weight of damages. None of it will be any simpler if contextual determinants point otherwise. Nowadays the SARS-CoV-2 pandemic is disclosing the actual extent of those determinants <sup>23,24</sup>. It featured firstly by procrastination albeit each handful days should have tallied, – right after Dr. Li Wenliang whistled for a new disease among his patients in Wuhan city, China by early December 2019 <sup>25</sup>. Secondly ensued that a month later, Chinese authorities issued incomplete information on severity, transmissibility and case fatality. And then, to make matters worse, an additional month lagged until the WHO Health Emergencies Programme pronounced the highest alarm level statement. Post hoc, the scenario of world havoc that stems from a local cluster seems predictable quite from its arrival on, serving as an oxymoron to the mandate that underlies the guidance of equity and health in all policies <sup>26,27</sup>. At the end of this dissertation, the social costs of the pandemic have not yet been quantified. By the same token, the control of chronic-nonreturn epidemics, granted their subtle manner of taking their toll, are subject to same contravening conditionals that very much hinge on the distribution of power, wealth and resources.



**Figure 1. Annual amount of IARC Group 1 carcinogen emissions in the proximity of 8098 Spanish towns (2007-2010).**

## 1.2. Cluster and clustering definition

Cluster has been conceptualized as *a rise* in the variability according to which cases tend to group together in the realm under consideration. A vital statistician should approach the



realm of the study to a stochastic process, from where clustering represents an ‘irregular’ grouping of cases/events <sup>28,29</sup>. Epidemiologically, a main concern about cluster definition stands on the marker within which variation (or variability, or fluctuation -used interchangeably-) occurs; namely time, space or other markers such as occupation, common-treatment patients, consumers of defective commodities, and so forth <sup>3,30</sup>.

This dissertation focuses on time clustering, i.e., cases gathered by shortening average separation at a given data set <sup>31</sup>.

Particularly I concentrate on retrospective, finite sets of event data. To be precise, two scenarios can be drawn: a) Public health departments addressing themselves to reports from persons concerned about disease occurrence in their place of work, vicinity or school, and b) Scrutinizing secular trends of chronic diseases in which legitimate interest leads to unearth clustering so far uncharted. Detection of clustering engages the process leading to elucidation of risk factors. Causative constituents are presupposed to aggregate in a similar pattern <sup>3,30</sup>.

And as with entitlement to monitoring utility, a prime interest in those clustering analyses is to obtain rapid alert of a higher incidence rate of the disease in the study population or population groups.

In difference from usual ‘false alarm’ occurrence, a genuine, true alarm temporal pattern is expected to include a cluster of the events. Besides, it is the irregularity of the time lag when recording what spotlights chronic diseases. This warrants methods here so that such natural irregularities not influence detection.

## 1.3. Nature of temporal chronic disease clusters

What follows is a schematic trial to delineate special attributes of temporal clusters. Certainly, several subtitles overlap. And many of the specifics listed or secondary features interest the flavour of exploratory testing, if others belong to a level subsequent to the latter.

Next lines focus on clusters of malignancies; although much of the notions would also apply in settings of other no-return diseases. This term articulates maladies that can be determined and reported *as is* only once for a given individual. They happen void of reappearances, irrespective of whether the morbid process of interest is curable or not, and may or not result in decease.

### 1.3.1. The low frequency attribute

Regardless of whether the disease(s) at issue is common or rare; clusters are realms, in which the malady *aggregates along limited* space, group, and time setting. Herein and perforce, whenever we run up against morbidity or mortality, aggregation would amount to small absolute count – gauged as low occurrence of observed cases.

Studying the incidence of chronic / no return occupational illnesses reifies the captioned connotation. Most of job-associated exposures occur indeed with low frequencies. The incidence of any putatively related disease will be low, even if a wide sector of the population is susceptible to develop the disease because of a given exposure.

Quite often, public health agents are asked for dealing with perceived (alleged) clusters. In a community, alarms light up as a result of the rise in a serious illness, usually cancer disease. By nature, most of these signs amount to frequencies distinctive of rare-diseases. For the nature of post factum generated data, the scientific method mandates the delimitation of knowledgeable-meaningful community stratum or geographic unit. Besides, ascertaining by rigorous and highest reliable case definition, depreciates the count of diagnoses per assessment. As an example, the US Federal Agency for Toxic Substances and Disease Registry (**ATSDR**) conducted an investigation of birth defects and Childhood-LHs clusters at the Marine Corps



Base Camp Lejeune, North Carolina. Inhabitants had been exposed to contaminated drinking water from 1950s to 1985. The participation rate was approximately 76%. Survey participants reported 106 cases among 12,598 children. Of these victims, 35 were diagnoses of neural tube defects (NTD), 42 of oral clefts and 29 of childhood-LHs. The ATSDR made extensive efforts to obtain medical information to confirm such reported cases and was able to confirm 15 NTDs, 24 oral clefts, and 13 cancers. Thus, in aid of validity, about a half of originally alleged cases became finally eligible for the research <sup>32</sup>.

Limiting the scope of cluster investigation to cases with refined diagnoses using fine-graded characteristics might confer a gain in causality research and policy. For either genetic or epigenetic traits, great expectations lie on their yielding towards fine-tuning, provided they lessen the misclassification of outcome <sup>8,33</sup>. Short-listing cancer or other chronic diseases cases by definite ‘omics’-dictated attributes should imply, however, fewer records to deal with upon cluster exploration <sup>34</sup>.

### 1.3.2. Actual or spurious incidence variation

Quite a small fraction of reported clusters warrants statistic examination. Clustering could be merely a result of heterogeneity in background population density <sup>35</sup>. Case-togetherness may well elicit suspicion of clustering, but it is worth seeking in unison the pattern of all people or all person-time at concerned category <sup>3</sup>. K. Rothman’s legendary lecture in 1990, points up the axiom simply <<. . . even people who do not get disease tend to be ‘situated close together’ in various ways. . .>> <sup>3</sup>.

Up to a sixth or a fourth of all the suspected clusters endure formal statistical testing <sup>36,37</sup>. Were the count of observed cases above the expected -given the size, age and gender of the concerned population- chance would continue to be a plausible interpretation for many of them.

Disease frequencies fluctuate, variation in the incidence rate is a feature of singular chronic disease diagnoses. We are seldom committed to account for broad group of tumours altogether; in most cases, investigation is focused on circumscribed and distinctive tumour

subsets and in a restricted population sector during a relatively short time span. The latter should hinge on in-built data set attributes, for example, data obsolescence; or else be contingent on characteristics of the population at issue, as in cancer clusters among infants or in limited occupational cohorts <sup>38</sup>. Nevertheless, it should be kept in mind that large random fluctuations around the mean operate on such sources of disease infrequencies.

Spurious incidence oscillations could also occur, should case ascertainment practices or registration completeness fluctuate. In addition, unmeasured or unknown factors may infringe the assumption of independence of observation. Cases occurring on a given time  $t$  may be correlated with the case occurrence on time  $t-1$ ,  $t-2$ , etc., as a result of *other than* causative factors <sup>14</sup>.

### 1.3.3. Appropriateness of event data and denominators

Prime cluster inquiry implies the measure of a rise from expected number of cases. Under the aforementioned unavoidable low frequencies, the quest warrants the choice of an appropriate reference population. Namely, to demonstrate an unusual occurrence, ‘appropriateness’, points up the provision of a “*good grasp of the usual*”; accounting for “*some measure of reality*” well ahead of any statistical analysis <sup>28,39</sup>. To begin with, actual strata distribution for main variables should be available – gender and age at least. Moreover, contemporaneousness makes a reference population suitable for the task too. The timelier, the more ready its representativeness.

Nevertheless, denominators provided to derive rates conform with census timing, and annual data rely on estimates for inter-censal years. Noisy data and potential instability on derived rates are treated through parametric, non-parametric and semi-parametric procedures. If no systematic bias is introduced by the interpolation methods, major errors will not be made <sup>38</sup>.

Besides, population-based registries contain the addresses of cases at the time of diagnosis. When high migration prevails, obviously the addresses may not be appropriate to explore clustering locally. That is, if the risk intensities are substantially influenced by study population turnover or by a changing prevalence of risk factors.



As far as paediatric cancer is concerned, more stable denominators ease the investigation of geographic clustering. Containing precise place of birth, counts of births register yearly. Clustering of childhood cancer in relation to place of birth can be readily laid on computer record linkage. On temporal clustering investigations, the linkage targets to infants born in a defined time period. Nevertheless, families of young children move house relatively frequently, thus, the address at diagnosis, or even the more informative place of birth, may mislead the actual place of causal insult <sup>35,38</sup>.

### 1.3.4. False positive and false negative errors and clusters

As regards statistical testing, cluster's nature exhibit distinctive sources of inflated alpha and inflated beta <sup>40-42</sup>. I use the term *efficiency* to denote the likelihood of a test in detecting an elevated disease incidence.

#### 1.3.4.1. False negative errors:

Most cancers occur in a frequency well below an annual rate of one per thousand. In order for a rise on these diseases to attain 'statistical significance' there must be a substantial relative risk. Malignancies with an annual rate of about one in 100,000, in a town of 5000 needs a relative risk as substantial as 8 to achieve significance with  $P$  value  $<0.01$ . Over a decade (let us suppose), this relative risk corresponds to 4 observed and 0.5 expected diagnoses <sup>43</sup>.

Efficiency of analyses is likely to be poor because of presumably multifactorial causation, carcinogens' transiency, low disease rates, unobservable induction and latency periods <sup>3,14,40,43-46</sup>. For instance, *post hoc* time limits (as though imposed by an investigation of an alleged cluster) – they may mismatch the actual period of the agglomeration in the series. Were an outbreak of morbidity to register over a given spell, those limits would possibly result in a gap between the epidemic and the observational stretch; thus tilting in the no-clustering hypothesis favor.

#### 1.3.4.2. False positive errors:

Inherent low efficiency comes coupled with an inbuilt-inflated significance that evolves from the huge and unknown sample space. Conceivable, rejection of the null hypothesis -of *no* increased incidence- may be provoked by so many no-return diseases, periods, communities, theoretical other endpoints that could have been tested. By assembling exploratory analyses on validated diagnoses and study windows carefully settled *in advance*, informed formal experts wipe off many of those inference biases <sup>47</sup>.

*Post hoc* time and region boundaries and age groups submitted to analysis, act augmenting the overall false detection rate <sup>36,40,42,48,49</sup>. In no instance has separation of chance effects at *a posteriori* situations been void of analytical difficulty. So much so that Coory and Jordan have advised avoiding the use of *P* values in clusters' investigations <sup>48</sup>.

In our general setting, methods able to detect material clusters, whether based on significance tests or not, might elicit many spurious ones <sup>18,31,35,39,50</sup>. On the other hand, since all cases of disease have causes, let us deem after Rothman (1990) <sup>3</sup> that phrasing <<*the cases are causally unrelated*>> inform more loyally of a cluster so called ascribable to chance.

Transcending the mere technical topic, note that every time people safety or wellness is at issue, strain supervenes between the burden of the false negative and false positive results upon the public – scathingly, about results' publishing or not thereof. Unavoidable so, I address myself to it, revolving around cancer epidemiology (see blue text box in [Appendix 2](#)).

All in all, decision making on temporal clustering should be rely on either most conservative scenarios, providing false positive concerns do exist or towering costs of any major intervention would hardly alleviate risk <sup>51,52</sup> or less conservative scenarios when precautionary principle policy or intervention, and false negative are at issue <sup>53,54</sup>. Obviously, cautionary practice and scientific research are not antithetical <sup>55</sup>.



### 1.3.5. The One-off feature

An inherent inflated  $\beta$  error can appear because of the low sensitivity of alternating research to replicate a material cluster. Every result of falsifiable concurrence of a disease cluster with an exposure cluster is susceptible of  $\alpha$  error, an effect of no repeatability in the population.

A failure to comprehend or to reproduce a material cluster does not preclude its validity. Distinctive causative paths, time-dependent determinants and other unobservable factors can uniquely condition the time aggregation of the concerned health outcome. The well-known clustering of childhood-LHs at a key nuclear fuel recycling plant called Sellafield in West England illustrates the point; the alleged series jointly unearthed by the community and television. The British government appointed an independent inquiry, which provided the confirmation of genuine cases as well as an excess of LHs cases within a particular rural district. Further, longitudinal studies established a clear relationship between male-prezygotic occupational exposure to ionizing radiation and high LHs incidence <sup>56,57</sup>. The linkage between parental exposure and leukaemogenesis in the offspring was subsequently discouraged by some authors revisiting the general topic of ionizing radiation in different exposure settings, and the health experience of residents under 25 years of age living near nuclear sites <sup>58</sup>. Arguably, any study conceived to address the particular occupational association should run up against a dissimilar, *unrepeatable* radiation exposure history. With this sort of statements, some researchers claim ought to be supported by convincing scientific explanations <sup>35</sup>. Except that accessibility to so called *explanations* should be subjected to postponement; an acknowledgment the author learnt (as a premed) after the Irwin D. J. Bross' tough letter to the editor of Environmental Health Perspectives by 1977 <sup>59</sup>. Back to our example, thereafter, not only have authorities given full recognition of the plausibility of transgenerational effects as a result of exposures to ionizing radiation, but research has also elucidated bedrocks and mechanistic pathways <sup>60-62</sup>.



Next, cancer epidemiologists use the construct of susceptibility. This domain highlights a subsector of the population *in fact* harmed due to exposures at sub-threshold levels; namely, exposure settings at which most people health is not influenced. Susceptibility behaves operatively either as a time-dependent (i.e. covering critical time windows of vulnerability) or time-independent factor. Time clustering may not be reproducible; for the causative setting exhausts the potential groups of susceptible people once upon a timely framework<sup>63</sup>. This, too, happens if susceptibility prevalence varies over time. Critical time windows are common concerns in reproductive and developmental outcomes but cancer<sup>32,64</sup>.

Brackets of susceptible people can suffer acceleration of disease incidence; for example, via mechanisms of exposure action strengthening the aggressiveness of noxious agents<sup>65,66</sup>. These individuals with exceptionally high risk are frequently unidentified, since most epidemiological data will not uncover accelerated morbidity experience nor do the studies reliably discern small relative risks<sup>14,67</sup>.

Earlier interventions may in several ways explain why the effects of health hazards in the immediate environment of individuals cannot be *pos hoc* determined. Conceivably, the causative agents in a population had been decreased before the onset of their systematic investigation, making them difficult to detect. Lack of health-hazard monitoring has precluded a confident assessment of past exposures<sup>2,68</sup>. So many times interventions coincidentally mend complex exposure milieus. All these real-life-scenarios hinder the reconstruction of either direct or indirect morbid response paths.

### **1.3.6. Clusters embedding in the series**

Tests of clustering aim at detecting an embedded aggregate within a given data set. Any single method may not be efficient in grasping the diverse temporal patterns of clustering. Clusters either may engulf all the data or bury themselves inside it; the outflow in the latter remaining unnoticed.

A temporal aggregate of cases within multiyear data set typically evolves from a causative factor newly introduced in the population.



Unobtrusive circumstances of enhanced chronic diseases incidence declines naturally. It is usual for a cluster span to entangle susceptible individuals in up to saturation. That is, until the malady harms all susceptible individuals. Embodied as ‘self limited’ epidemics, constitutes rather an issue of relatively closed populations, e.g. a workplace community. Concerning the earlier tail of the series, reasons include the lead-time to diagnosis and that only a small fraction of the community responds to the exposure. Such issues contribute to blur the depiction of the departure time from the baseline incidence rate.

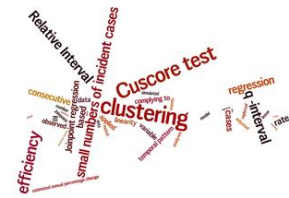
Substantial clusters are able to surrogate an otherwise ambiguous secular incidence trend or even a soothed one thereby depicting concurrent person-periods of dissimilar risk levels<sup>40</sup>. Exposures or ill proneness may boost a natural carcinogenic process in a less conspicuous and sparsely fashion. These can encompass a medical deleterious procedure or treatment, an environmental accident that causes a new and selective contamination source in the community, or exposure of production workers to an emergent somatic-cell mutagenic agent. An epidemic cancer by itself may endow a causative path for an embedded cluster. This is indeed the widely known phenomena of second primary malignancies, regardless of treatment (so that the pathobiology timing is undetectable), yielding a more subtle or sparse clustering over the time. Second primary solid tumours as well as leukaemias follow radiation therapy in childhood by reportedly swelled range of latency periods<sup>69,70</sup>. With second primary leukaemia, a complex *reciprocal* relationship exist with thyroid cancer, a disease with annals of classic outbreaks.<sup>71</sup> Similarly, enhanced risk of the second primary leukaemias evolves from either systemic (no radiotherapy) or other shared proneness to breast cancer<sup>72,73</sup>. The path modality though, holds for other chronic health outcomes: birth defects clusters can result from damaged DNA in sperm cells from pre-treatment cancer patients<sup>74</sup>.

Of the diagnoses, it is also possible that the latter or the earlier ones disclose aggregation, whereas the overall count of cases is close to or even beneath prediction. Reckon with the well-known residual confounding termed *healthy worker effect*, or *the healthy mother effect*. Such should be the case in the following illustrative scenarios: Bove et al. (2014), upon

assessing Standardized Mortality Ratios (SMRs) for all cancers, some cancers of primary concern and non-cancers among marines and navy personnel chronically exposed to contaminated drinking water found a value  $<1$ <sup>75</sup>. These SMRs, calculated using the US death rates as the standard population, led the count of cases in the exposed population to fall below prediction. But using Cox-extended regression models, the exposed persons showed elevated hazard ratios for meaningful degenerative causes of death<sup>75</sup>. The healthy mother (or father) effect should be anticipated anytime that the incidence of non-return health outcomes in childbearing parents is lower than that of comparable adults at equal age bracket<sup>76</sup>. Women and men who desire *and preserve* the capability to conceive children do represent a relatively healthy sector among the adult population<sup>77,78</sup>. Contemplate now a sequential inquiry on a community of adults on their reproductive ages, who secularly second the postponement of gestation. Suppose the concerned chronic disease outcome of interest is caused by a time (in fact, age)-dependent as well as a built-up of childbearing-associated pathogeny (e.g. osteoporosis). And let us circumscribe the study population on couples that have been rearing  $>4$  offspring. Conceivably, a cluster will be triggered only at the earlier years of the series – unless the chosen population of reference shows a comparable endured health profile<sup>79</sup>. In this particular sequence and an especially imposing trend of postponing parenthood, the first years of the time series show the highest outcome intensity, but, in contrast, the excess of cases may no longer be apparent in the last years.

#### 1.4. Techniques for temporal cluster inquiries

The underlying featured attributes of chronic disease clusters drive the implementation of techniques inspired on epidemiological reasoning. To date, we have walked through the realization that temporal clusterings of infrequent-no-return diseases seldom evince themselves <<as a trout in the milk>> – paraphrasing XIX century American philosopher H. D. Thoreau (1850) – to connote a circumstantial evidence strong as it stands<sup>80</sup>. Ahead I expound on



exploratory sequential tools, whose tailored properties satisfy current benchmarks to approach such demanding issues <sup>11</sup>.

#### 1.4.1. Sequential techniques for temporal cluster inquiries

Pinpointing huddles of any infrequent-no-return malady on the time axis can seldom be met because the causative circumstances and the period from insult to diagnosis are unknown. Moreover, inquiries should be related or unrelated to a suspected exposure.

Cluster identification misleads by simply staring at a huddle of events if population heterogeneity is not taken into account. In most of the real instances, the study responds to a concerned community. Analyses are then applied to a retrospective data set that does not include a control group. Proper control groups could well not be available or relevant.

As far as commencement measurements are concerned, investigating clusters of infrequent diseases warrants indicators with detecting properties that point estimates like the standardized incidence ratio (**SIR**) <sup>81,82</sup> will not outreach. This centuries-old ratio exhibits downsides; including but not limited to: 1) The difficulties checking off specific rate proportionalities between a reference and a given study population <sup>83</sup>; 2) The handling of denominator data not truly representing person-time schedules at risk <sup>14,44,84,85</sup>; 3) The instability of the SIR upon small -indeed up to 0- mean incidence number of events <sup>63</sup>; 4) The discrete property of its random variable (periodic observed number of cases); or 5) By engrossing a whole data set, its failure to articulate the increased rate trend over stretches of a given span <sup>86</sup>.

Models fail to grasp the full range of biomedical real-life possibilities because they are too strict <sup>28</sup>. As implied above, rate measurements for low frequency outcomes could be of the zero value. Models conceived of narrowing vulnerability to the called *zero-inflated observations* of the explained variable ought to be assessed for fitting to the data. This is a major issue of alternative model-seeking mostly related to the realm of space clustering <sup>87,88</sup>. Besides, reported upgrading of detection statistics accompany intensive computational effort; see for instance Cançado et al. (2014) <sup>89</sup>.

*Sequential* type methods are useful when we do not posit a suspected causative agent. In any case, they account for not knowing beforehand when the rate has been starting to increase<sup>14</sup>. Analyses of clusters eager for sequential schemes on cumulative data and not just, for example, annual examination of the events<sup>84</sup>. Sequential tests aim at avoiding delays in the detection of the elevated rate.

Sequential stands for accumulation of results of consecutive analyses. Each analysis carried out at the end of each year or subsequent to the appearance of each group of predefined number of cases. In this context, THE SETS, the Cumulative score (CUSCORE), the LIKELIHOOD RATIO TEST, the SCAN, and the Cumulative Sum-like techniques (CUSUM) are all sequential methods.

Clustering surveillance refers to sequential techniques applied to ongoing data, over an undefined period. As the analyses are performed repeatedly, a significant result is bound to occur however much stable the morbidity is. That is, a meaningless ‘significance’ in such a context. Thus, an inbuilt disadvantage of routine monitoring consists of too many false alarms<sup>29,41</sup>.

At this point, it is worth highlighting pertinent attributes about some of the monitoring-based techniques. The Cumulative Sum-like techniques (CUSUM), are highly sensitive in the manufacturing process, and typically operate under an array of monitoring<sup>90-93</sup>. CUSUM technique systems have been servicing the surveillance of congenital malformations after 1960’s- Thalidomide disaster<sup>94,95</sup>. The SETS was used and suggested by the European Economic Community for surveillance of birth defects in small communities<sup>96</sup>.

With CUSUM, the appraisal of events happens at fixed time intervals; hence neither calendar independence nor identical time-distribution of counts can be assumed for certain kind of health processes<sup>29</sup>. Ultimately, whereas low efficiency has been proved for CUSUM for rare-health-events, it efficiently detects large rate increases<sup>95,97,98</sup>.



The Cumulative Score (the CUSCORE) , as coined by Munford (1980), encompasses the monitoring of a normal distributed variable to elicit corrective action in the control of continuous production processes <sup>99</sup>. Chen (1978) <sup>94</sup> suggested THE SETS technique that lies on the time intervals between consecutive cases using the *RELATIVE INTERVAL* (**RI**) statistic. And Wolter (1987) <sup>92</sup> executed the *RI* statistic into the CUSCORE. These two procedures aim to detect elevations of intensities when the yearly-count of cases is scarce.

In fairness to the actual scope, let us establish a margin between monitoring sequential procedures for clustering surveillance and significance tests for clustering applied to *a given series* – to skip the former hereinafter.

THE SETS and the CUSCORE alarm-signalling tests of significance have come after Chen & Goldbourn (1998) <sup>100</sup>. The relative efficiency of these two tests differs over the range of increased rates. In relation to the envisaged scenarios of low level of increased rates, the CUSCORE's efficiency (either as a test or as a monitoring system) surpasses that of the former <sup>92,100</sup>.

The original CUSCORE operates whenever each event appears. Chen (2001) <sup>101</sup> extended both the surveillance and the test of significance to be used as the waiting time until the *r-th* case is observed, where *r* is a predetermined positive integer.

Assuming independent time-appearance of the cases, such waiting time follows the Gamma distribution; the distribution of the time-interval between single diagnoses is exponential <sup>102,103</sup>. As a matter of fact, chronic diseases, as well as most no-return health outcomes fit in with the independence assumption. In spite of this, some dependency between diagnoses may occur, i.e., when diagnosis activities are for some reasons enhanced, driving individuals to medical checks. Of malignancies, that remains valid, but it does not in infectious diseases and other non-infectious endpoints with arguable 'contagion' path – as in the phenomenon of temporal correlation between time units in reported suicide clusters <sup>83,104,105</sup>.

Current knowledge supports the multifactorial genesis of no-return diseases and unequal risks over population subgroups. The multifactoriality notion leads to consider alternative statistical underlying distributions of the data set. In respect of this, Wolter (1987)<sup>92</sup> claims that even under a log linear trend, CUSCORE is superior to THE SETS. This versatility for the task of cluster identification at underlying distributions different from the simply exponential is not superfluous.

Occasionally, time and/or space distances in tests of clustering are subject to mathematical models that provide risk adjustment. These can modify the statistical power. Attributes of entered constants effect the outcome of the significance too – see authoritative quotations on these issues at the appendix of CDC’s Guidelines for Investigating Clusters of Health Events,<sup>106</sup>. Except the CUSCORE -as well as THE SETS- test works as a counterexample for if it does not pivot on mathematical transformations to variables. The statistics it is based on do not bare constants added to actual time distances, nor does it admit any arbitrary cut points. Hence, the statistical power of the CUSCORE test, hinges only upon the data setting as it stands, whatever it be<sup>100</sup>.

Consequently, the CUSCORE is clearly more efficient to tackle subdued increments of incidence and if risk adjustment is paramount. As a side point, the CUSCORE outperforms the Joinpoint regression regularly used in prediction of cancer incidence<sup>107</sup>.

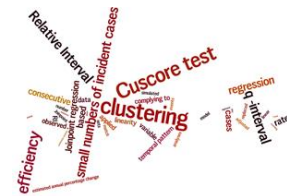
#### **1.4.2. Playing characters**

##### **1.4.2.1. The awaited r event**

The  $r$  event denotes a fixed positive integer. The size of each event  $r$  equals the number of cases per event. In general,  $r$  amounts the upper integer of the annual expected incident number at baseline.

##### **1.4.2.2. The relative interval**

Focusing on the assessment of time until alarm upon monitoring rare events Chen (1978)<sup>94</sup>, suggested an interval statistics that accounts for the expected number of the events’



set. Then, this hub statistics was used to demonstrate a surveillance system for chronic diseases based on time interval between consecutive diagnoses <sup>107</sup>. The notation  $RI$  grew coined latter by N. Mantel (Dr. Chen -personal communication- 2018).

The  $RI$  statistic reflects the waiting time to event. One can consider it stable under temporal changes in the risk of the population because it is -as explained below- a standardized statistics that controls for size and profile differences between populations <sup>40,86,92,100</sup>.

$RI$  is the ratio between the observed and the null expected interval  $E(w)$ , which is calculated according to known baseline rates. Thus,  $RI$  is constantly the expected number of cases *whatever* the current  $E(w)$  length is. As such it has a gamma distribution and its mean is the event size  $r$  <sup>108</sup> (where  $r$  is the number of cases in each event). Note that the exponential distribution conforms a special case of gamma when  $r=1$ .

The expected time until event equals the event size. This is so since the expected time until a single case is  $E(W)$  months, hence the expected time until  $r$  cases is  $r * E(W)$ , thus  $E(RI) = r * E(W) / E(W) = r$ .

Coming back to the advantages of  $RI$ , let us pay attention that if  $r$  cases are expected annually, the relative interval reflects the reciprocal of annual  $SIR$  <sup>63,107</sup>. However, the advantage of  $RI$  over a periodic  $SIR$ , such as a subperiod ratio between observed and expected number of cases, stems from the fact that  $RI$  is a *continuous* random variable whereas, let us remind,  $SIR$  is based on a *discrete* random variable (i.e., number of observed cases).

Measuring the time interval in standard units rather than in plain months enables the *control* of relevant changes in the studied community. Translating an observed interval into  $RIs$  warrants assessing the expected number of events in each year under study – if need be (if the expected incidence is changing). Therefore  $E(w)$  adapts outright to the current expected incidence. By definition, **one unit** constantly amounts the period during which one event is



expected. The unit may vary according to time, disease or from one town to another. Yet it expresses the expected time for one event <sup>63</sup>.

The *Relative interval* comprises both a time and a count measure, inasmuch as each unit in *RI* is the expected time between events, it, represents the number of cases that are expected during the observed (waited) time gap.

An example showing this double notion of *RI* follows:

The expected number of cases among *N* individuals in *year i* is  $\lambda_i = 0.86$  (or 1 event in 14 months)

An event was observed after  $W_i = 7$  months. The number of cases  $x_i$  expected in 7 months is:

$$E(x_i) = (0.86/12) * 7 = 0.502$$

$$RI = W_i / E(w_i) = 7/14 = 0.502$$

An event registered in an interval whilst merely 0.502 should have been expected.

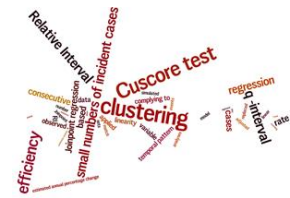
Under stable risk (i.e. under the null hypothesis), *RI* is expected to equal 1 and a smaller value whenever elevated risks register. Therefore, *RI*s shorter than expected reflect elevated disease rates.

#### 1.4.2.3. The *q-interval* (*qi*) derived from *RI*

The *q-interval* was postulated by Chen in the mid 90's and first applied to analyses of registered data of several leukaemias in a city in Israel. <sup>40,109</sup>.

The distribution of the time-interval *q* between consecutive diagnoses is exponential; so be  $r=1$ ,

$$qi = \exp(-RI)$$



Recall that  $RI$  is tailored to the conditions prevailing during each year, namely, a control framed by tailoring the definition of the time unit according to the null expected number of diagnoses. Temporal changes in size and structure of the community are controlled by updating the definition of time unit. Thus, the inferred  $qi$  is a standardized statistic.

The  $qi$  statistic is defined as the a priori probability that the  $r^{\text{th}}$  case is diagnosed *after* the actual observed date, or it is the null probability for fewer than  $r$  cases (e.g. that not more than  $r-1$  cases) register within the observed period. In turn, this last sentence can be restated as: ‘during an interval in which  $RI$  cases should be expected’ – because with  $RI$  the calendar observed period is translated into the expected number of diagnosis in such time interval. Since the event size amounts to  $r$ , this equally matches the enunciation that  $qi$  represents the probability for no event within each attendant inter-event gap. From the standpoint of the time-span,  $qi$  equals the probability than an interval is longer than that observed. Yet again it can be stated as the probability for a longer  $RI$  than that observed inasmuch as one expresses the interval in terms of  $RI$ . From now on these definitions are used interchangeably.

Although  $qi$  is evaluated as a probability, it is actually a random variable, since its value is determined by a random time interval ( $RI$ ). As cumulative continuous distribution, Its distribution is uniform over 0-1 and its expected value  $E(qi)$  is 0.5<sup>110</sup>. When the cases are more frequent than expected, as happens within a cluster, the  $qi$ -values are likely to be greater than 0.5<sup>42</sup>. The consistency of large  $qi$ -values over several events as well as the position on the observed series should render clues for interpretations.

There remains an accompanying feature of the  $qi$ : the convenience of disentangling hidden veers in incidence intensity. When the rate increases at some unknown point of time one can expressly reckon a subsidiary approximation based on the median  $qi$  denoted  $q_m$ , from which one shall supplementarily refine the intensity ( $\gamma$ ) estimate<sup>42</sup>.

$$\gamma = \frac{\ln(0.5)}{\ln(q_m)}$$

### 1.4.3. The alarm-signalling test. The cuscore's basics

The CUSCORE relies on defining each observed  $RI$  as either 'short' or 'long'. It is defined as 'short' if  $RI \leq RI_{crit}$ .  $RI_{crit}$  stands for critical values. Under baseline conditions,  $p_{crit}$  represents the null probability that the observed  $RI$  be shorter than a critical reference value termed critical RI ( $RI_{crit}$ )

$$P_{crit} = Pr(RI \leq RI_{crit})$$

These critical values are presented in the *-appended-* reference 97 Table 1 therein, according to  $r$  and  $s$  number of events. Following a short  $RI$ , the CUSCORE increases by 1; otherwise it decreases by 1. The score recycles (i.e. equals 0) if it is negative. The *alarm* implies a significant result, coinciding with a score=5. The  $RI_{crit}$  values ensure significance at *one-sided* 5% level (for details refer to Chen & Goldbourt -1998-<sup>100</sup>). I shall keep using the term *alarm* under such connotation throughout.

### 1.4.4. Confirmatory procedure

Confirmatory techniques deal with the potential of high chance of false alarms. Hence, they target on the reassurance of a prior positive result in the sequence.

There have been 2 confirmatory techniques proposed by Chen et al. (1993) <sup>41</sup>. Both are nurtured through data from the period of time succeeding to the alarm. Specifically, using the preceding  $RI$  for each of the first five diagnoses made subsequent to an alarm. These post-alarm tests techniques aim at rejecting 75% of false significant results.

### 1.4.5. Graphical display by accumulative q-interval ( $sq$ )

So far, the described attributes of  $qi$  frame its usefulness in graphical display of the temporal changes of the incidence rate <sup>42</sup>. As aforementioned,  $qi$  is expected to be 0.5 under stable conditions and larger under elevated rates. However, provided the small counts of cases and a moderate increasing trend induced by a new aetiology, the increasing  $q$ -intervals may not offer a demonstrative graph as the *accumulative q-intervals* (denoted  $sq$ ) would. The resulting curve is sensitive to small elevation in rate, because the effects of any elevation (on  $qi$ )



accumulate. Herein, one should note that the expected  $sq$  value amounts the ordinal number of the assessed interval divided by 2. The slope of the  $sq$  curve equals 0.5. Interpretation of the  $sq$  curve, however, bears nuances <sup>42,107</sup>.

Statistical evidence is only one ingredient for us to tackle decision-making in cluster reports <sup>49</sup>. Persistently, a non-statistically significant overall test should not preclude the investigation of an apparent individual cluster. <sup>35</sup>. Confirmatory tests by themselves plausibly fail to accept a genuine cluster if, for instance, vulnerable individuals -to a definite path of causation- run out in a small community with low turnover, for instance, a factory <sup>63</sup>.

The empowering of a graph as a starting point in an epidemiological investigation associated with cancer or other chronic diseases deserves an underscore. Unlike false alarms and in the long run, the temporal pattern of a true alarm is expected to include a cluster of the events. The temporal pattern of the events also provides hints for interpretation and may guide towards a more exhaustive research. At any rate, one wishes to see this depict reflecting on whether and when about an elevation of rates has begun.

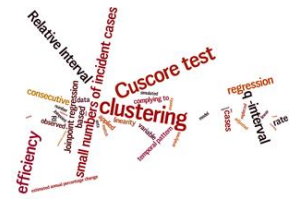
Through these last subsections, I have pinpointed the outstanding procedure tools. A primary attribute of these tools is their matching a *sense-the-data* aspiration. This implying analyses of the sequential data ‘as is,’ in concordance with the criteria used by health agencies statements to choose a statistical cluster method – see for instance from CDC’s guidelines (2013) <sup>11</sup>. Furthermore, as just expounded, the standardized interval measures are *interlinked with each diagnosis* upon tailoring the definition of the time unit according to the null expected number of diagnoses. In this context, the controlling ascertainment takes into account the differences between communities or even by background variability in pathogenicity. Such qualities provide robust results by easing them readable, communicable as well as internally valid.

## 2. THE STUDY. THE ANCILLARY CONNECTIONS

Even though the sequential approach introduced thus far had been used for simulated data and various real-live prompts of perceived neoplasm clusters, hub testing with CUSCORE has been run at *pre-hoc* scenarios in Israel <sup>101</sup>, but seldom otherwise. This dissertation bounds to demonstrate an integrative approach that performs resourcefully by transcending the macro-level standardized rate trend descriptors: down to municipality and down to specific cancers whose frequencies pertain to the range of rare diseases (or less than 1 in 1600-2000 citizens) <sup>111,112</sup>.

The case for action stems from an acknowledged trend of increasing incidence for each of three Myeloid Malignancies categories in Girona province during a period of 15 years 1994-2008 <sup>113</sup>. Using raw data from the province-based cancer registry, the authors pronounced a seeming incidence rate steadiness of Acute Myeloid Leukaemias (**AMLs**), as well as significant increased rates of Myelodysplastic syndromes (**MDS**), Myeloproliferative (**MPN**) and Myelodysplastic/Myeloproliferative neoplasms (**MDS–MPN**).

A driving grasp of MMs and the so far explained waiting time techniques shall complement each other. Accordingly, next paragraphs serve the purpose of informing interpretation of the results, priming oneself for deciphering supervening arrays of temporal clusters as for these particular rare diseases. Of course, such an undertaking should not void an assimilation of the up-to-date cause-wise morbidity. The closing subsection on nomenclature summarises physiopathological advances. This being managed by carefully meshing the WHO's conceptions upon the eve of the actual study's data (by 4<sup>th</sup> edition 2008 -based on the 2000' International classification of diseases for oncology (3<sup>rd</sup> edition) [**ICD-O-3-**] with the subsequent and latest ones (the agency's classification of those tumours revised 4<sup>th</sup> edition 2017) – to service same goals thereof <sup>114–116</sup>.



## 2.1 Clusters of lymphohaematopoietic malignancies in communities

Published clusters on haematolymphoid neoplasias in the literature dates back to mid-1900s. Next three quotations have been selected on purpose for they concern to particular stretches of the lifespan; not least these previous studies yield breakthrough knowledge on these rare-diseases. In 1972, Vianna et al. documented a one-off infectious pattern for Hodgkin’s disease in New York State lasting 20 years. Most of the cases diagnosed among non-consanguineous young adults <sup>117</sup>. Authors’ postulate on infectious cause, however subjected to several challenges, drags into the acceptance of likely exogenous etiologic components for these lymphomas either <sup>118,119</sup>. Antenatal exposure emerged as a plausible explanation after in-depth investigations revealed cluster of leukaemia in children born in the Village around Sellafield nuclear facility (West Cumbria, England) – yet not among those who moved to the area <sup>14,56,120</sup>. A third exemplificative study focused on Myelodysplastic syndromes (MDS) – a rare MMs’ subset known to peak among the elderly. Namely, spatial clusters spring in a large north-eastern area of USA. Covariate data on biomedical variables and personal follow up had been achievable for Liu et al. (2015) <sup>121</sup> who came up with insights on epidemiologic features, prognosis and aetiology. For example, the result of geographic distribution, which, since it did not arise associated with the biological aggressiveness of the disease, led the authors to postulate a thesis on the lack of association with point sources of exposure.

By doing environmental research on cancer, whenever this entails longer induction and latency periods, population mobility deserves attention. Amid the fact that complete residential history barely exists at population-based registries, the haematopoietic malignancies short latency periods (which peak at 5-10 years) confer certainty on the geographic profiles of these malignancies’ outcomes <sup>8</sup>. This would not hold for several lymphohaematopoietic and solid tumours. And it is worth noting that families bringing children up move more often than other occupants; thus, the foregoing certainty may not work the same in the case of paediatric LHs

## 2.2 Myeloid Malignancies

The present section incorporates current features of MMs by starting with background biomedical facts that should assist with our investigational insights. I just addressed ‘*omics*’-dictated diagnoses selection apropos of the low-frequency attribute of temporal cluster of chronic diseases -see [subsection 1.3.1](#). This conduit, fostering the hope that specificity of outcome would operate to the advantage of causation inquiries.

If only a joint case-ascertainment of dissimilar maladies pivoted on ‘*omics*’-dictated commonalities improve statistical power <sup>123,124</sup>. Not least that it ultimately repay effective policies of prevention. An example in furtherance of this goal, comes from molecular epidemiology research, featuring common mutants of isocitrate dehydrogenases (essential enzymes in epigenetic regulation and several cellular processes), and implicating a mixture of solid tumours *along with* MMs <sup>125,126</sup>.

Commonalities features within and between the main MM categories render a working standpoint for us to look at these malignancies; and thereby permitting us decipherable results with the postulated research questions. In heading for the actual realm, acquired or inherited susceptibility, kindred proneness, as well as exogenous causative agents constitute a basis for inferring or providing for the elucidation of small-communities’ chronic disease epidemics.

And in-depth genomic characterization spins with the update on knowledge of the natural history of myeloid malignancies. As well as it sheds light on certain distinctive characteristics by subtypes. In unison, underlying pathophysiological processes concurrent with karyotype damage and / or driving mutations are being disentangled <sup>116,127–129</sup>. It ought to be of interest to our realm provided disruptions’ characterizations end up informing the construct of a continuum from normality to haematological preleukaemic state through chronic phase, and to progression to transformation /aggressive-full-blown disease <sup>130</sup>. Were taxonomic entities of a main category to follow a thread through this rationale, one should envisage a profit in understanding leukaemogenic clustering.



### 2.2.1 Commonalities

Since all these myeloid lineage disorders originate from altered pluripotential hematopoietic stem cells, it is not surprising that divisions between these tumours are irregularly faint <sup>116</sup>. Broader indeed, the crosswise lineage plasticity across LHs is likely unfindable amongst other cancers. The following paragraphs summarise some of the held in common among the four main groups of MMs in the following paragraphs:

Myeloid malignancies tend to evolve over time from less to more aggressive forms of disease. Both MPN and MDS, *transform* into AML. Moreover, transformation among different cell lineages is a current event (e.g. from myeloid to lymphoid LHs) <sup>114,116,131,132</sup>.

Alongside transformations, still consensus taxonomy does bare revisions, up to the pledge of repositioning of entities to *a distinct* MM category. See for instance how mounting knowledge about confluent clinical characteristics makes of a particular acute myeloid leukaemia imitating a myelodysplastic syndrome <sup>116,133,134</sup>. Mastocytosis, formerly regarded as an MPN has been shifted to a detached category at the revised 4<sup>th</sup> WHO classification <sup>116</sup>.

Besides, *common* abnormalities have already been determined to a quite broad range of LHs. Two examples should suffice: the translocated *ABL1* (v-abl Abelson murine leukaemia viral oncogene homolog 1) – that is, the *BCR-ABL1* fusion-gen contained in the Philadelphia chromosome (**Ph+**) <sup>135</sup>; the *JAK2V617F* (the somatic point mutation on gene Janus kinase 2 (**JAK2**)) <sup>116</sup>.

Except for monoclonal gammopathies discriminating MMs from Lymphoid-allied neoplasms, particular gammopathies can precede *many of* the latter but *seemingly* no one of the former <sup>136</sup>.

Upstream commonalities are also being revealed by the application of next generation sequencing <sup>137</sup>. A novel leukaemogenic somatic mutation in *PRPF8* gene associated to AMLs, MDS and MDS–MPN, represents an illustration <sup>138</sup>.



Other loci commonalities for upstream-upregulation have been reported: the ecotropic viral integration site 1 (Evi1), an overexpressed gen in leukaemia cells, as compared to normal T lymphocytes, is an example. This comprising defects of molecular components that control haematopoiesis and entail poor outcomes in AMLs, MDS and a prime MPN-entity <sup>139,140</sup>.

As far as downstream aftermaths are concerned, shared features deserve quotation too. It should be exemplificative that prognostic commonalities among allo-transplanted patients operate on both MDS and AML. Here, residual disease detectable by flow cytometry, CD3 chimerism has consolidated. Actually, by behaving as a factor for relapse after allo-transplantation at late post-transplant period, and predicting overall survival <sup>141,142</sup>.

Side by side, pathways of damage partaking of unlike carcinogenic agents add an extra aspect, insofar as a variety of MM diseases are involved. Benzene and cytotoxic drugs testify to it <sup>143</sup>.

### 2.2.2 Constitutional proneness and acquired susceptibility

Heritable mutations of germ cells, as well as particular chromosomal disorders are unequivocally related to incremental lifetime predisposition of specific leukaemias in the context of clinically heterogeneous syndromes <sup>116,144–146</sup>.

Age of onset sometimes interlinks with susceptibility instances. Documented proneness to anticipation phenomena (determining progressively earlier ages of manifestation) for certain familial MMs has been inventoried by Johns Hopkins University's OMIM – see for example entry # 601626 therein, on familial acute myelogenous leukaemia <sup>147</sup>.

But inherited genetic risk, granted, may not have sufficient-independent bearing on MM causation. Proneness has been shown for endorsed environmentally sensitive genes. For example, concurring with a risk to confirmed MPN subtypes that present the foregoing JAK2V617F clonal marker -such as Polycythaemia Vera (**PV**), Primary Myelofibrosis (**PMF**), and Essential Thrombocythaemia (**ET**)- <sup>33</sup>.



Of course an increased incidence among next of kin leads to be interpreted as inherited, but shared environment or cultural transmission of hazards among family members always deserve to be sought and pondered <sup>148</sup>. Undoubtedly, <<environment>> does not preclude exposure to haemato-oncological treatment of relatives. This holds true even though correlation between disease occurrence rises with genetic proximity <sup>149–152</sup>.

The defensible assumption of a past generation *acquiring* tumour susceptibility traits renders diverted by paradigms like those exaggerating the weight of genetic variation or chance ('*bad luck*') <sup>153</sup>. Thence therefore, acquirability of either somatic or germline hereditary carcinogenic distortions renders undersized regarding aetiology. Needless to say, such credence does not bear criticism from a standpoint of current knowledge in epidemiology <sup>154–158</sup>.

Furthermore, pre-zygotic exposures related to increased risk for LHs have been reported <sup>56,159,160</sup>. These and other malignant tumours have been recorded either for first progeny or beyond the first post-exposure generation in rodents <sup>60,161,162</sup>.

### 2.2.3 Exogenous causes

Exogenous causes of MMs cannot be understated. This is of an understatement, in proportion to the cumulative knowledge entitling current consensus – despite the growing milieu of exposure complexity that leads to designs of observational studies on human cancer overwhelmingly demanding or by nature non-sensitive. Extrapolations from their results or from nonhuman models -regardless of the robustness of their outcomes or the weight of mechanistic evidence- wait untenable lags, sometimes generations, until the preventive measures serve the public.

Concordance for LHs in monozygotic twins is high for the early life yet it does not hold through later life; a hint of prevailing environmental causes <sup>163</sup>. Interestingly, an established heritable proneness can take years to manifest. But until then, the newborn and infant haematopoiesis appears to be grossly healthy. For example, individuals with mutated GATA-binding protein 2 (*GATA2*) have an increased lifetime risk of myelodysplastic/acute myeloid

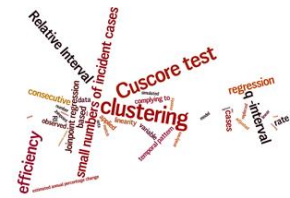
leukaemia<sup>146</sup>. Nonetheless, a weak penetrance is typical, and the affected individuals experience an indolent progression<sup>146</sup>. And researches interpret that other causes, including environmental may be decisive to the development of disease<sup>164</sup>.

Studying international cancer registry series lead to realize less regional variation among youngest brackets. This fits the indication of prevailing determination by causes other than inherited / genetic proneness in LHs. That is what exactly propose Mendizabal et al.(2016)<sup>165</sup>, using the series of chronic myelogenous leukaemia (**CML**) from the ‘Cancer Incidence in Five Continents’ to describe a definite paucity of inter-regional variation among those diagnosed before age 50. In contrast, a prominent and significant variation exists when the median age at diagnosis is  $\geq 50$ , both genders<sup>165</sup>.

Certain carcinogenic settings, may wreak havoc the worst at older people because of a postponed time of effect over the induction gap. Such effect modification by age has been recalled so as to explain an exceptional susceptibility of older workers to benzene’s leukaemogenicity<sup>166</sup>. Bearing comparison, older but not younger sick persons experienced an excess of AML subsequent to exposure to external-beam radiotherapy for female breast cancer in a cohort from the National Cancer Institute’s Surveillance, Epidemiology, and End Results database<sup>73</sup>. Similarly, increased age at autologous stem-cell transplant is an independent risk factor for therapy-related MDS. And so has been quoted revolving around AML as an aftermath of primary cancers<sup>61,143</sup>.

Younger people, however, are, in general, reckoned as more susceptible to develop cancer after exposure to exogenous agents<sup>72,73,168–171</sup>. And adult cancers occurring at unusual younger ages has been marshalled to flesh out a presumptive patent exogenous -mostly preventable- induction as well as of accelerated carcinogenesis<sup>170,172–177</sup>.

Turning now to gender variation, a recognized fact is that main LHs, as most cancers do, have male gender predominance. Causal interpretations arise whenever the association with gender does not coincide with the anticipated one. In this sense, observational epidemiology abides by the dictum of the economy of explanations. And leading Occam’s razor-guided



hypotheses seldom put forward constitutional risk factors but pathways of exogenous determinants <sup>178</sup>. Consider for instance therapy-related MMs, which depict a straightforward sex-specific incidence rate ratios < 1, and its parsimonious attribution to the iatrogenic effect of breast cancer treatments <sup>73</sup>.

Ionizing radiation exposures at diverse settings cause LHs <sup>10</sup>. After irradiation at low levels, the hematopoietic bone marrow will develop cancer above the ‘spontaneous’ baseline rate. <sup>179</sup>. And will arise with an excess greater than that recognized for other tissues and organs with proneness to radiation cancer such as breast, gonads, or thyroid. Furthermore, the cumulative evidence is such a compelling as to question current diagnostic and medical surveillance benchmarks on behalf of patient safety. <sup>169,180</sup>.

Solely or in conjunction with cytotoxic drugs, radiotherapy predisposes to various LHs <sup>10,72,84,114,116</sup>. Thyroid carcinoma patients treated with radioactive iodine (<sup>131</sup>I) at therapeutic doses as low as 999 megabecquerels have been reported to develop CML and AML. In fact, two-logarithms under the elsewhere-published doses given to cases that subsequently developed Acute LHs (i.e. tens of thousands of megabecquerels) <sup>181,182</sup>.

Cytotoxic and accompanying therapeutic adjuvant drugs have been associated with leukaemogenesis and cytogenetic subtypes described so far for MDS, AML and MDS–MPN <sup>10,114,116,143</sup>. Regarding MPN, the causal relationship is not yet well established <sup>183</sup>.

Benzene was postulated as haematotoxic more than a century ago. Furthermore, this xenobiotic has been proven leukaemogenic, indeed unsafe at any magnitude of exposure <sup>10,184</sup>. Mechanistic links to leukaemia and accruing observational support on this prevailing organic solvent are continuously within scope <sup>143,185</sup>.

Another existing domestic and occupational contaminant with recognized genotoxic and mutagenic potential, formaldehyde, has been classified by the International Agency for Research on Cancer (**IARC**) as a cause of MMs with sufficient evidence on humans <sup>10</sup>. Reviewers of this determination have since challenged this statement in some respects <sup>186,187</sup>.

Finally, leukaemogenicity is convincingly associated with mixed activities or milieus and compound carcinogens<sup>10</sup>. The blend of xenobiotics regarded as <<Smoking>> counts as a recognized cause for MDS<sup>121,167,188</sup>. Presumptive adverse association between occupational exposure to biocides and risk to be crippled with AML has been recently found by a meta-analysis of longitudinal studies<sup>189</sup>. Moreover, remarkable sources inventory occupational scenarios immanent of particular job titles or industry branches<sup>190</sup>. For instance, the <<boot and shoe manufacture and repair>>; <<the styrene-butadiene industry>> -here studies show a dose-response relationship between excess of leukaemia and cumulative exposure to butadiene alone; whereas studies from this monomer industry (without checking on styrene's or other exposure covariates) show an excess of LHs in general-<sup>191</sup>.

### 2.3 Nomenclature and leukaemogenic clustering

Eventually, upon assessing epidemics, public health insights could be enriched when scrutinizing the main MM diagnoses of the clustering cases: thereby tantalizing a collateral outreach out of the pervasive classification fundamentals; despite their rather therapeutic and prognostic orientation – so far being patterned amidst the rapid-increasing information on pathogenic genes.

In the wake of incoming information on mechanistic pathways, fresh knowledge has been consolidated on the pathobiology of LHs. Therefore, irreversible carcinogenicity is usually associated to the *accumulation of* driving somatic mutations, complex karyotypes ( $\geq 3$  cytogenetic anomalies) and parallel continuous changes -e.g., aberrations and mutations succeeding genetic instability;- actually correlating strongly with dismal prognosis<sup>116,127</sup>.

The myelodysplastic syndromes are comprehended among the *mature*, as opposed to *precursor-immature*, MM varieties<sup>192</sup>. This handful of entities taxonomically durable and characterized by paced and heterogeneous clinical conducts, thus lends itself to conceptualising a ladder for perniciousness -or risk and prognosis grading. Still morphologic and cytogenetic



metrics underlie a widely endorsed *risk* array. Actually the system also awaits ongoing enlightenments to incorporate knowledge from accruing somatic mutations <sup>116,167</sup> – so far as ‘hits’ of driver mutations are assumed to presage the MDS distinguishing dysplastic clonal expansion and disease outcomes <sup>130</sup>.

Explicitly, discrimination has been tailored among MDS entities along an ordinal continuum; to pose chronically disarrangements at critical pathways of biological consequences, by a parallel takeover of increasing risk for life-threatening illness. Three discernible levels arise <sup>116,167</sup>: 1) the first level consists of a *low risk* group represented by 3 subtypes: refractory anaemia (renamed MDS with single lineage dysplasia, code ICD-O-3 9980); MDS with ring sideroblasts and single lineage dysplasia; and MDS with isolated deletion (5q) – codes 9982 and 9986 respectively; 2) the second level applies *mid-risk* to MDS with multilineage dysplasia (formerly refractory cytopenia with multilineage dysplasia) -code 9985;- and 3) the third or *high-risk* level, which consists of MDS with excess of blasts (formerly refractory anaemia with excess blasts) -code 9983. It should be kept in mind that this scheme does not include the unclassifiable MDS (ICD-O-3 9989), due its clinical heterogeneity.

Even so, the revised WHO’S fourth revised classification patterns blood-cancers on no gold standard <sup>116</sup>. And the rest of MMs categories just lack an analogous gradient as delimited by the severity of the insult by subtype. Think of MPNs, too, of a *mature*-variety MMs: either the archetype CML or the others, show up as exhaustive subtypes for a cancer registry’s nomenclator; i.e. each rendering the connotation of whichever severity phase.

Following suit MPNs, and likewise included among the *mature* MMs, the MDS–MPN category presents subtypes whose defining features do not inform about a discriminating intensity of carcinogenic damage among them. As a hint, the survival of CMML patients (which comprise nine tenths of the MDS–MPN ones) ranges between 2 and more than 100 months <sup>193</sup>.

All AML subtypes, the prototype *immature-precursor* neoplasms, encompass in addition to structural chromosomal rearrangements, multistep acquired mutations in  $\geq 3$

biological pathways – over epigenetic regulation, transduction signalling, transcriptional regulation, that underlie processes like cell adhesion, apoptosis inhibition, transformation, proliferation, invasion and metastasis <sup>116,194,195</sup>. Interacting nuances determine diverse prognoses and mosaic remissions, as well as relapses and drug resistance as per discrete therapeutic arsenals <sup>196</sup> – beyond the scope of this research. All the same, AML subtypes converge around their -too shared- overwhelming clinical course. Yet knowledgeable authorities resort to counting *separately* therapy-related myeloid neoplasms (ICD-O-3 9920 & 9987) from the remaining *de novo* AML or other MMs counterparts <sup>116</sup>. This plea, albeit proceeding from a dominant if legitimate desire to clarify biomedical variants further, is worth transferring towards enquiries on leukaemogenic clustering. Finally, expert cancer networks are considering that many among the AML with multilineage dysplasia subtype (9895), lately denominated AML with myelodysplasia-related changes, lie in the confines between AML and MDS <sup>116,197</sup>.



### 3. RESEARCH QUESTIONS AND OBJECTIVES

#### 3.1 Research questions

The recognition of 15-years upward trends for myeloid-lineage LHs in an entire province merits further inquiry <sup>113</sup>. This is a challenging obligation, eased by the high quality and completeness of the population-based data underlying such observations. The 4 MM main categories comprise homogeneous diseases, meaning so an homogeneous group of cases to start with.

Here ought one to steer the inquiry plan that to unveil causative clues, one begins by asking according to what distribution of time and space those rare disease cases had occurred. A not very perceptible incidence rate could reflect on the inherited type of cancer, whereas larger increases argue for material environmental effects <sup>198</sup>.

Noteworthy, the observed trend could have started after 1994 and/or ended before 2008. It is reasonable to assume that if there was causative cluster for an increasing trend, an equilibrium state should be reached at a certain time. Else, a still increasing trend must be related to a recent cause or to a recent intensification of a causative cluster. Should one fail to uncover a worsening risk status for malignancy for any larger stretch, a temporal cluster of the events still could be modestly gestating. Since one aims at reacting upon these instances, one appeals to analytical tools whose performance grows in parallel to the accumulation of events.

Results will indicate if the reported increased rates likely impacted on each municipality in the Girona province or just some of them. Contemporariness of distinguishable indicative disease clusters at same place points to strengthening interval validity thereof. One should expect to obtain some indications regarding the time of initiation of the increased rates. A timeline revealing clusters of transforming-prone MM categories (like MDS or MPN) that



launch earlier than clusters of the end stage MM categories (particularly AML), provide further plausibility for a just preceding and then pervasive causative cancer mechanism in a community.

The results of the analyses could well provide clues regarding the plausible nature of the carcinogen(s) suspected to be involved <sup>107</sup>. Thus, for example, if clustering arises in most or in all the studied municipalities, a relatively new prevalent single cancer mechanism may be suspected. Else, be the interpretation, a medical practice-enhanced diagnoses. On the other hand, be the increase observed in just some of the municipalities, common attributable factors would be hypothesised for the communities in such localities.

Of cancers with peak incidence from plain adulthood on, a clear-cut gender deviation from expectancy incites to interpretations on cause-relatedness. Except reproductive cancer sites and the marginal issue of constitutional susceptibility, a gender differential in cancer morbidity points to a distinct exposure to exogenous determinants, inherent to a gender-selective factor; commonly, but not limited to, a work-related one <sup>178</sup>.

Besides, patent shifts to an unusual age bracket at diagnosis, generally underpin acceleration of carcinogenesis or the reckoning of a driving induction for malignancy contingent to a prevalent acquired vulnerability <sup>177</sup>. A deviation from predicted presentation age entitles a justification to proceed from an exploratory level of investigation up to a more intensive one<sup>199</sup>.

In addition, fine-grained scrutiny by morphology subtypes among MMs categories over clustering cases may be auxiliary to the insight on the nature of carcinogenicity. For instance, should the AML and MDS therapy-related entities (ICD-O-3 code 9920/3; 9987/3) take over a particular aggregate of cases, frequent iatrogenic (chemotherapy and/or radiotherapy) source-like is to be hypothesised toward next incidence analyses. Else, be in-cluster prominence of subjects fundamentally expressing worst disarranged karyotypes, one is to contemplate either persistent or insidious leukaemogenic exposure clustering.



This is a **non-ex post investigation**. Given the exploratory flavour of these analyses, no falsifiable hypotheses on clustering are being set, nor is there any progressive consideration with respect to any postulated path of common causality.

## 3.2 Objectives

The objectives of this dissertation are:

- To assess for temporal clustering of MMs through waiting time techniques. The layout progresses through the four main MM categories -AML, MDS, MPN and MDS-MPN- and municipalities in the province of Girona over a 15-year period. When needed and data is available, confirmatory procedures are operated to lower false alarm.
- To explore the intensity, time pattern and overall place-dependent behaviour of any detected cluster.
- To analyse person-data to help inform possible causative paths on the communities bearing indicative clustering. That is, per gender, age and beforehand defined morphological subtypes of concerned primary interest; their deviation from the usual –for which *unusual* is evaluated according to a reference population.

## 4. MATERIALS AND METHODS

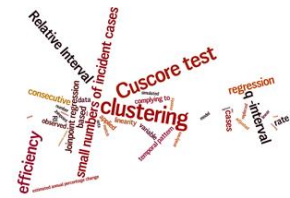
### 4.1 Outline

Core procedures were as follows: First, calculation of overall *SIR observed:expected* (O/E) ratios. A value of one indicates that the observed is equal to that of the expected count for a particular municipality for the period. Second, any time these achieved a figure  $\geq 1.0$ , but not significantly so, I examined the waiting times sequentially and CUSCORE-tested for time clustering alarm. Third, confirmation by wiping false positive out on subsequent intervals where available. Fourth, plotting chronological patterns of the standardized interval statistics (observed and expected). Fifth, univariate examination person variables for each exposed huddle.

For every main MM-category, the approach was checked off across preset communities in the Girona Province: the analyses were restricted to all municipalities that registered at least 10 cases for any of the four MM categories i.e., AML, MDS, MPN, and MDS–MPN. Therefore, the study consisted of analyses with prespecified boundaries of time, persons and place. The reference population was predetermined too. Fifteen *municipalities* (best affordable specifics that stood as a proxy for *communities* or *towns* – whence I interchangeably used the three terms) were so preselected.

The maxim about the improper use of, the confidence intervals, the P values and power calculations has been heeded<sup>200–202</sup>. The sequential test, although cardinal in my analyses of temporal clustering, is inspected under scrutiny. It has never been used as an exhaustive source of judgement. Instead, the quality of the data, constituents such as the understanding of underlying mechanisms, the integrative contribution from the evidence in its entirety are privileged over statistical measures (e.g. *P* values or confidence intervals when applicable).

Cases of myeloid malignancies diagnosed from 1 January 1994 to 31 December 2008 were analysed. Crude rates per 100,000 inhabitants/year were 3.60 for AML, 5.18 for MDS, 5.53 for MPN, and 0.74 for MDS–MPN.



## 4.2 Reference Population

The standard population was extracted as an aggregate of the annual counts over the whole of Girona province (731,864 people in 2008). The demographic distribution of which was obtained from the Institut d’Estadística de Catalunya (IDESCAT). Accurate population profile by age and gender was only available for the period 1999-2008, as explained below in [Section 4.4](#).

## 4.3 Myeloid Malignancy cases

Diagnosed cases of MMs were extracted from the population-based *Registre de Càncer de Girona (GCR)* in 2013. Of them, the registry held a 96.3% completeness of ascertainment at its catchment area, and a joined rate of histological verification of 98.6%<sup>203</sup>. The GCR fostered outreach retrospectively to ensure the complete coverage of these diseases<sup>113</sup>. Codes were harmonised following the criteria established in the third edition of the International Classification of Diseases for Oncology 2000 (ICD-O-3)<sup>114</sup>. In addition, comparable standardized incidence rates tallied with the population-based registries from abroad<sup>113</sup>. A non-negligible percentage of 27.7% of MDS patients received diagnosis as unclassifiable subtypes (code ICD-O-3 9989). By the taxonomical state-of-the-art, the *unclassifiable attribution* relies on a close follow up – till *final* diagnosis becomes feasible (see the revised 4th edition of WHO’s Classification of tumours of LH)<sup>116</sup>; in consequence, this was taken into account by *excluding* ICD-O-3 9989 cases, upon computing subtype distributions inside clustering of MDS cases. This is in adherence to the call made by this thesis’ contemporary guidelines too. That is to say, to excluding any diagnosis-based <<*unknown*>> from calculations of percentages of cases -criteria synopsised by HAEMACARE Working group (2010)<sup>114</sup>. Likewise, upon seeking into AML clusters, computations on relative frequencies of *de novo* subtypes set aside the

multilineage dysplasia (ICD-O-3 9895) and the therapy-related subtypes (9920 & 9987) –see rationale in [Section 2.3](#).

Time, place and person data for each case included age at diagnosis, gender, date of diagnosis (accurate to day), subtype (a morphologic MM-entity code), and community of residence (i.e., a postcode for each of the 221 province’s municipalities) at the time of diagnosis.

#### 4.3.1 Descriptive analyses on person data

##### 4.3.1.1 Age and Gender

For a given MM category, communities with indicative cluster (labelled ‘*clustered community*’), were assessed for by-gender overall *observed:expected* -O/E- ratios, and median age at diagnosis from one end to the other, each along the entire sequence -which is assumed mostly ascribable to the clustering.

##### 4.3.1.2 Morphological subtypes

Frequencies of confirmed malignancies’ subtypes involved in a cluster were calculated. Specifically, the period prevalence, and the ratio of observed to expected incidence during such spell. In order to anchor chronologically a start point for adding thereof, a simple a priori rule of thumb was set. Be an example and the associated rationale as follows.

Refer to Table 6 showing an indicative imbedded cluster of MPN. Notice that each event involves 4 cases. Here the clinical diagnosis marking the 7<sup>th</sup> event dated on the 26<sup>th</sup> of March 2003, only 6.3 months after the previous event, the 6<sup>th</sup>. For the actual purpose, all the diagnoses so registered from 1 January 2003 onwards, would be factored in. Thus for the sake of ad hoc counts of MM-entities or subtypes let us reckon any diagnosis that is recorded from the *first day of the calendar year* in which the clustering has just arisen, as if they belonged in. A workable ‘cut-off’ point, granted because under clustering, the transition to a worse risk status likely transcends whatsoever diagnosis charting date.



As far as any inclusion of  $P$  values is concerned, it served informative purposes only. Nor are statistical significant testings being used to interpret those findings.

#### 4.3.2 Subtypes' prevalences

Amid a clustered community, a subtype (or allied subtypes) predominance along the clustering period was defined: when inside a cluster span, the prevalence of a concerned primary interest subtype(s) out of its pertaining MM-category cases became larger than the corresponding proportion among *<<clustering-skipped communities>>* named too *<<non-clustered communities>>*; explicitly, those shortlisted localities that were spared from any MM cluster, as ought to have revealed the implemented sequential procedures.

I termed *<<concerned-ancillary subtype(s)>>* (or *subtype(s)* or *entity(s)*, for shortening) a beforehand chosen disease. This selection was made insofar as an entity or an allied subgroup was endorsed as a settled (steady) illness by the updated WHO nomenclature, and whose diagnosis would ever be deemed final. The specifically excluded entities are listed at the opening of the current Section -*Myeloid malignancy cases*. The *ancillary* quality was defined if, according to current knowledge, variations in the occurrence rate service the interpretation whenever a temporal indicative cluster has been detected. Feasibility of meaningful interpretations imposed per nature the grouping of subtypes. The author is responsible for the final log of those concerned primary interest entities, thanking the senior haematopathologist Ms. Natalia Lloveras (MD) for the reassuring assistance, and priming oneself by reviewing up-to-date references –see [Section 2.3](#). For the stipulated breakdown of subtypes, see [Subsection 5.7](#).

#### 4.4 Expected morbidity

The expected number of cases was calculated for each of the four MMs categories, using the specific incidence rate as per 36 age and gender groups in Girona province all through

1994-2008. Denoting by  $R_s$  the age and gender specific rate in the reference population and by  $N_{i,s,t}$  the relevant group size in each municipality  $i$  in year  $t$ , the expected number of new cases in year  $t$  is given by:

$$E(x_{i,t}) = \sum_S R_s * N_{i,s,t}$$

The sum of these values over the fifteen-year period represented the expected number of diagnoses per municipality and disease category. Age and sex structure was unavailable regarding the first lustrum of the period. Therefore, for the period 1994-1998, the expected yearly figure was estimated as five times the nearest accessible annual value (1999).

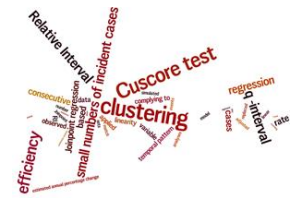
The assumed baseline incidence rates could indeed deviate from the true baseline values. This deviation could stem from random error in the estimated specific rates. In that instance, it would be reasonable to assume positive deviations for some specific rates and negative for others. Since the expected number of events within each interval amounts the sum (over all individuals at risk) of those specific rates, the impact of these deviations is expected to more or less nullify each other. Therefore, any effect resulting from so little biased baseline is likely to grow ineffectual.

#### 4.4.1 Test of significance for SIRs

Standardized incidence ratio over the 15 years (1994-2008) was evaluated and the observed number of diagnoses tested for significance using the Poisson distribution assumption, with means equal to the expected morbidity counts. In this approach a one sided  $P$  value is called for to target a one-sided question; that all or most of the data are clustered (attendant one-sided  $P$  value  $< 0.05$ ). No cluster test was accomplished on data showing significant  $SIR$  results.

## 4.5 Observed interval as random variable and RI appraisals

The time interval (in months) was evaluated for each consecutive group of  $r$  events. The size of the  $RI$  was calculated by an MS-Excel function allotting years of 360 days. For a



particular sequence, whenever 2 diagnoses registered at the same date, an obliged = 1 day interval value was assigned.

For each studied community the first waiting time departs 1 January 1994 (the very first day of the data collected), ending at the observed day of the first group of  $r$  events. The second waiting time starts at the date of this  $r$ -th event and ends at the date of the second group of  $r$  events. Of course, under increased rates  $r$  events do occur in a shorter time. As a rule,  $r$  amounts the upper integer of the annual expected incident number at baseline  $E(x_i)$ .

Following that,  $RIs$  were evaluated for each interval.

Each  $RI$  be defined as either “short” or “long”. A “short” interval indicates increased rate. An  $RI$  will be defined as short when smaller than or equal to  $RI_{crit}$ , a *critical value*. Critical  $RI$  values are available in published tables for  $r = 1$ <sup>63,86</sup> and for  $r > 1$  in the [appended article – Table 1](#) therein<sup>107</sup>. These critical values had been calculated such that the null probability that by chance we get CUSCORE = 5 within  $S$  intervals be 0.05.

For example, if number of consecutive intervals analyzed ( $S$ )=10, and  $r=2$  -namely the data include 20 cases. An interval will be defined as short if  $RI \leq 1.311$ . This implies that 2 cases registered during a time when not more than 1.311 would be expected.

#### 4.5.1 Tailoring RIs

Many of the studied municipalities were sparsely populated. And, by the mid of the present times series, every second one of them inhabited by not more than 800 residents<sup>204</sup>. Also, starting by mid-90’s, population movements took place in and out of the study communities. Abiding by the unobservable potential influence of unstable age-gender structure on our estimates, somewhat attributed to migration, correction was imposed to our data series: rather than inferring  $RIs$  assuming a constant value of  $E(x)$  for the entire set of data, the figure was updated every 4<sup>th</sup> year whenever achievable:

1.  $E(x)$  of 1999 was used on intervals for years 1994-1998, and years 1999-2002. Thenceforth,



2. 2003's, and 2007's  $E(x)$  were used on intervals for years 2003-2006 and 2007-2008 respectively.

This updating aimed at reducing the effect of a possible moderate trend in the reference population.  $RI$  values, however, can grow biased each time they do not stem from adequate reference data, yielding expected intervals between false alarms. Related to this, it is worth mentioning that an effect upon signalling of a 5% bias in the estimated baseline rate has been shown as reasonably moderate by Chen et al. (1997)<sup>205</sup>.

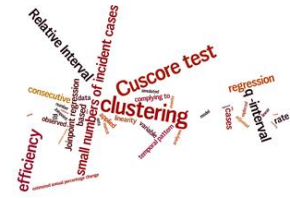
#### 4.5.2 The alarm signalling technique

The CUSCORE test is based on a score accumulated over sequential analyses. A decision must be made upon the advent of each event. Starting at 0, the cumulative score is either increased or decreased by one after each analysis. The score is increased by 1 when  $RI \leq RI_{crit}$  and decreased by 1 otherwise. When the CUSCORE is negative, it is assumed that the data observed that far do not indicate an increased rate and accumulation of the score restarts. As can be realised, this control chart endows with the ability to observe each event by itself, which means an advantage over the overall count of events. Yet another property makes this test, by design capable of alarm signalling for any stretch within the analysed data, the more remarkable.

Results are significant when  $CUSCORE = 5$ . It takes purposely that the overall false detection rate be 5% (a one-sided). This should be, for an incidental clustering accounted *at any place* within the analyzed data.

Terms and procedures could be presented in a simple hypothetical example:

The data included 10 cases diagnosed over 14 years since January 1 2008. The CUSCORE analysis in that series would be based on the time interval between consecutive cases (namely,  $r=1$ ); hence,  $S=10$  interval assessments. Suppose that the first diagnosis occurred on 30 September 2008 and the second on 1 February 2009. Be 1.1 the assumed incidence rate for both 2008 and 2009. This gave an expected interval until an event of  $12/1.1$  months or 10.9; and  $RI=9/10.9=0.826$ . To be precise, 0.826 cases were expected during the observed time interval.



In short, so calculated  $RI$  would be the expected number of events during the observed interval. The value of  $RI_{crit}$  for analyses of 10 sequences using  $r = 1$  is 0.473. Consequently, the observed interval would be defined as “long” and CUSCORE would remain 0. The waited interval until the second case was 4 months and  $RI=4/10.9=0.367$ . This means that the second diagnosis would have been made over an interval during which only 0.367 diagnoses would have been expected under stable conditions, that is, less than the critical value (0.473). Hence, this interval should be evaluated as “short” and the CUSCORE after the second diagnosis would score 1.

Before proceeding, let us refer to a couple of caveats: firstly, the possibility of an accidental significant result by the CUSCORE test deserves a reminding. In no instance should one evade the need to confirm clustering by one way or another. Secondly, this test, as well as any other, may frequently mislead true clusters. It is important to note that except for the series of Girona town in this study, each  $RI$  pivoted on a single case. Thus, *long RIs* may frequently occur even under an elevated rate, not to mention the effect of randomness and uncharted cases. The results of a confirmatory test as well as the cumulative *sq* curve add by design an extra layer for the evaluation of a cluster as true or false.

#### 4.5.3 Handling incomplete recording of dates

As aforementioned, MMs cases over the whole province through the 15-year sequence had been rather completely registered by GCR (N=1331). The year of diagnosis was known for all the cases. By category type, missing data proportion for month of diagnosis fell below 5%; the greatest thereof (4.5%) corresponding to the MPN category –as could be expected because of its natural history. Uncharted day of diagnosis ensued in low proportion of subjects too (11%). Naturally for a given sequence of cases, missed exact date of diagnoses bring about inaccuracies of  $RI$  estimates.

December was the calendar month when the lowest count of diagnoses was registered for the 4 main MM subsets throughout the full period 1994-2008. This sort of administrative-unobservable lag, should at most defer an alarm, and hence operates against the sensitivity of the temporal clustering test.

A very important ramification of the procedure deserves mention: that an inaccurate date impacts the  $RI$  appraisal only if it borders on the associated  $RI_{crit}$ . Whenever the accurate month of diagnosis was missed for a count of diagnoses in a known calendar year, the possible interval arrays between the cases were exhaustively gauged. Since our series focus on rare events, such sort of lacking data was barely met. Those chronological variants have a bearing on the waiting times of those adjacent -contemporary- cases for whom more accurate dates are recorded. Upon encountering a milieu like that, the shortest and longest possible intervals and their  $RI$ s were derived out of each. If events with fully registered date took place in the same calendar year, their dates were used as anchors. Thereafter, a confident use should be allowed *if both* the minimum *and* the maximum possible event-associated intervals drove a *unique direction* judgment, i.e., “short” or “long”. Rather, one ought to interpret the full sequence with caution. Abiding by the same caveat, upon computing ensuing statistics required ad hoc for a descriptive temporal pattern (see in [subsection 1.4.2.3](#)), median  $RI$ s were adopted whenever imprecise dates elicited unsteady attendant  $RI$ s<sup>109</sup>.

If data lacked precise day of diagnosis a minimal mistake was imposed. Assuming diagnosis took place in the middle of the months (the 15<sup>th</sup> day), *minimum-maximum departure* from any factual interval was framed. The exceptional cases needing this calendar adjustment were underlined in Tables. As can be easily understood, an appreciable effect of the exact day on the CUSCORE result is very unlikely.

## 4.6 Censored intervals

The interval of the last event may actually be censored. The CUSCORE test is able to use that censored interval if the derived minimum possible  $RI$  value is *larger* than  $RI_{crit}$  in correspondence to  $r$  and  $S+1$  events. In such instance, the CUSCORE decreases by 1 (unless current value is zero, in which case, would stay zero as pre-set). Otherwise the use of a censored  $RI$  is prevented.



The rationale to permit an associated censored interval –that is longer than the critical value for  $S+1$ , relies on the realization that  $RI$  measured at  $S+1$  *already exceeds* the  $RI_{crit}$  for  $S+1$ . In such situations, any exact –incoming– date whatsoever, *would not* anymore add relevant information.

#### 4.7 Confirmatory procedures

An elicited alarm was accepted or rejected by applying the confirmatory approach of Chen et al. (1993)<sup>41</sup>. Explicitly, if a CUSCORE of 5 was achieved, five or sometimes less subsequent intervals to the alarm are further evaluated.

According to the preferred technique, either the median or the mean  $RI$  value was contrasted with reference intervals termed  $t1$  and  $t2$ ; as follows:

- a) The reference interval  $t1$ , inferred with regard to the percentage of false alarms one intends to eliminate; aimed at confirming just 25% of false alarms
- b) The reference interval  $t2$ , calculated with respect to the required power; set at a twofold rate rise, and aimed at rejecting 2.5% of true alarms.

The reference values of  $t1$  and  $t2$  associated with each of the techniques are based on the assumption that the time interval has a gamma distribution.

One confirms the alarm if the statistics (mean or median standardized interval) value is smaller than the corresponding critical reference value  $t1$ . One reject it if is above  $t2$ , and reserve judgment if the value is between  $t1$  and  $t2$ . The critical reference values are characteristic for each of the 2 methods.

The advantage of the median over the mean-based route is that one can sometimes confirm or reject the alarm before the fifth or even the fourth diagnosis. Be advised that among the three first observed  $RI$ s the largest value is the maximum possible median of 5 values, while

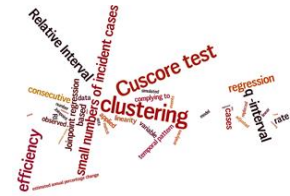
the smallest value is the minimum possible value for the median. Accordingly, one confirms the alarm if the maximum of the three *RIs* is shorter than  $t1$ , and reject it if the minimum of the 3 observed *RIs* is longer than  $t2$ . Moreover, one can occasionally reject an alarm even before the third diagnosis. This occurs once the two first *RIs* arise longer than  $t2$  and the time interval since the second diagnosis just evolves longer than  $t2$ . Thus, the median is useful whenever the number of postalarm diagnoses is fewer than 5. Otherwise the mean procedure should be used since its power is somewhat greater for the mean based technique <sup>41</sup>.

On the other hand, uneasiness for eventual missed cases also endorses a reason to resort to the median-based route – yet given the data present, it appeared rather counterfactual; owing to the nature of these diseases as well as the earnest charting by the Cancer Registry. Chen et al. (1993) <sup>41</sup> considered choosing the median over of the mean more appropriate in settings of low morbidity rates on a *monitoring* array. A “*not to wait for*” new recorded cases at the end tail of our series was adopted, following a timeliness-guided principle <sup>206</sup>: lest not fairly available data from the Cancer Registry cause unnecessary delay for confirming an alarm. Otherwise, the mean-based has been used.

The confirmatory tests enable rejection of 75% of false significant results. The attendant probability of confirming a true significant result depends on the increased rate <sup>41</sup>. The referenced intervals, under the mean based test are  $t1=\underline{0.6737}$ , and,  $t2=\underline{1.0242}$ . For exemplificative scenarios, Table 5 provides a deferment confirmation instance, and, Table 6, a sequence containing enough post-alarm observations and a clear-cut confirmation test.

## 4.8 Graphical display of temporal pattern

As said before,  $qi$  is expected to be 0.5 under stable conditions and greater than 0.5 under elevated rates. Cumulative  $qi$  was denoted as  $sq$  and used to present the temporal pattern of events <sup>42</sup>. In view of the small counts of cases and assuming a *moderately* increasing trend induced by a new exposure, the *sq-plots* shows up the rising  $q$ -intervals <sup>42,107</sup>. Notice that the



slope of the line could be misleading since the derivation of each  $sq$  depends on the preceding  $q$ -*intervals*. Indeed, the strength of this tool arises with the *accumulating* evidence. Besides, the curve might not grow accumulated quickly so to show high rates at the beginning of the series, even if they actually were. And, obviously, by virtue of these study settings, the  $sq$  curve could hardly include both the commencement and the termination of an epidemic activity.

In certain settings, as remarked apropos the alarm signalling technique, no significant result is obtained by the CUSCORE test, whereas the  $sq$  curve exhibits elevated slope. This is because it mirrors a sum of consecutive intervals and thus softens the impact of outliers. Had one failed to rule out missed cases or chance occurrences in the series, the curve would have looked as though composed of several slopes (as an example, see Santa Coloma de Farners' series on MPN – Table 3 and Figure 4).

Comparison of observed and expected  $sq$ -plots was accomplished either for any  $O/E \geq 1$  or whether the CUSCORE spells clustering or not.

Note that in line with its larger dataset, Girona town had a tally  $> 1$  for  $r$  values. So as to derive  $qi$  under the assumption of gamma distribution (rather than the exponential – where  $r$  equals the unity) one could use the Poisson <sup>101</sup>.

## 4.9 Software

The statistical analyses for quality control on morbidity data, and for the denominator data arrangement of age and gender, year by year schedule, relied on ‘*aggregate*’ and ‘*breakdown*’ each run on IBM® SPSS® Statistics version 23, 2015 (SPSS Inc., Chicago, IL, USA).

Sequential analyses were carried out on Microsoft Excel Professional Plus 2010 (Microsoft, USA). The period prevalence of primary interest MM-subtypes and, for its

precision, the 90% confident limits were computed under the binomial distribution (Wilson's Score) using an open-access calculator under the auspices of D. Eayres (2014)<sup>207,208</sup>. R 3.5.3-base package's functions were used to execute Monte Carlo simulation; also, the Odds Ratio tests were retrieved from its '*fisher.test*'-stats package<sup>209</sup>.

#### 4.10 Quality control

The crude dataset was issued by the Girona province's population-based cancer registry. The GCR data complied with the standards of IARC and so merited inclusion in 'Cancer Incidence in Five Continents' publications<sup>210-213</sup>.

Univariate analyses were applied on the whole raw data set of morbidity. Place, time and person variables as well as MMs subsets were *examined diagnostically*. Descriptive tables proceeded from syntaxes such as: `/cells=count; COMPUTE filter_$=( ); FILTER BY filter_$`. Both unlikely and missing values were annotated and inspected. Incomplete recording of dates was handled as previously described (see [subsection 4.5.3](#)).

#### 4.11 Bioethics

The cases data was delivered by GCR devoid of any personal identification of the cases, which is systematically kept confidentially and anonymously by the registry. In no instance has the author access to its pseudonymisation system. As to the research blueprint, it is of a non-interventional order either.





counties according to their descending number of communities showing either such excess of cases or scoring for temporal cluster by CUSCORE (this means the cumulative score reaches a 5), or, if attaining a graphical pattern (by the *sq-plots*) compatible with clustering, whichever. At last, I tidy each of these communities with positive clusters in keeping with a descending number of concurrent indicative clusters for any of the 4 MMs categories. Since this dissertation targets the introduced system of procedures, I shall refrain from exhibiting data not contributing nuances of utility.

## 5.1 La Selva County

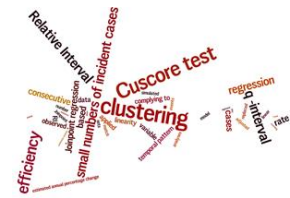
La Selva was the county with the greatest number of communities showing a likely cluster. Then, all 5 of the La Selva's shortlisted communities underwent sequential analysis as follows.

### 5.1.1 Santa Coloma de Farners

Santa Coloma de Farners is La Selva County's capital, with 70.6 Km<sup>2</sup> and a density of 162 inhabitants/Km<sup>2</sup> in 2008. This municipality was submitted to assessment for possible huddles in 3 categories: AML, MDS and MPN. Except for the latter, Poisson probabilities were  $< 0.05$  (Table 1). Table 2 depicts the expected numbers per MM category and the observed counts (underlined therein).

With the AML temporal cluster, Figure 2 displays the temporal pattern of expected vs. observed *sq* statistics. The curve of the cumulative observed *q-intervals* had a steeper than expected slope from the very second event (November 1997) embracing 9 cases.

Regardless of *RI* recorded, the by-gender O/E implied a greater excess among men (Table 10). The median age at diagnosis (63) was somewhat younger than their province's contemporaries (68). And within-cluster diagnoses were void of cytogenetic preponderance. Except one individual presenting multilineage dysplasia, all the rest encompassed 8 proven *de novo* AML subtypes. Its period prevalence hardly exceeded that of all the communities



devoid of clustering (88.9 vs. 83.0%) along the study period (Table 11). For the 11.04-year huddling, the expected figure for those *de novo* entities had been  $E(x)=3.004$ .

**Table 1 Cases per community-year and Observed:Expected ratios for 15 preselected municipalities 1994-2008.**

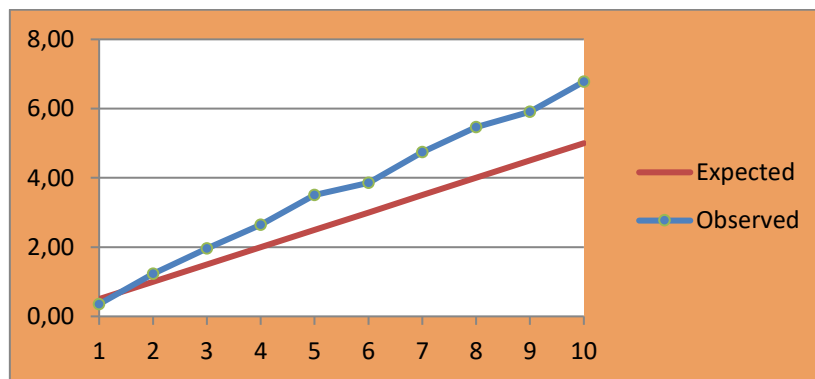
Disease [Cases community year]	Municipality (County)	O/E (Poisson P)
<b>AML</b> (N=151) [1.01]	Figueres (Alt Empordà)	0.54 -
	Sant Feliu de Guíxols (Baix Empordà)	0.99 -
	Palafrugell (Baix Empordà)	<b>1.55</b> (0.129)
	Olot (Garrotxa)	0.80 -
	Salt (Gironès)	0.99 -
	Girona (Gironès)	<b>1.08</b> ns
	Banyoles (Pla de l'Estany)	<b>1.42</b> (0.145)
	Lloret de Mar.(La Selva)	0.98 -
	Santa Coloma de Farners (La Selva)	<b>2.03</b> (0.029)
	Blanes (La Selva)	<b>1.21</b> ns
<b>MDS</b> (N=204) [1.24]	Figueres (Alt Empordà)	0.39 -
	Sant Feliu de Guíxols (Baix Empordà)	0.65 -
	Olot (Garrotxa)	0.80 -
	Girona (Gironès)	<b>1.04</b> ns
	Salt (Gironès)	<b>1.10</b> ns
	Cassà de la Selva (Gironès)	<b>1.83</b> (0.025)
	Banyoles (Pla de l'Estany)	<b>1.23</b> ns
	Lloret de Mar (La Selva)	<b>1.06</b> ns
	Santa Coloma de Farners (La Selva)	<b>1.73</b> (0.036)
	Blanes (La Selva)	0.82 -
Ripoll (Ripollès)	0.92 -	
<b>MPN</b> (N=260) [1.33]	Figueres (Alt Empordà)	0.92 -
	Sant Feliu de Guíxols (Baix Empordà)	0.77 -
	Palamos (Baix Empordà)	0.74 -
	Olot (Garrotxa)	0.89 -
	Salt (Gironès)	0.93 -
	Cassà de la Selva ) Gironès	<b>1.93</b> (0.017)
	Girona (Gironès)	<b>1.21</b> (0.059)
	Banyoles (Pla de l'Estany)	0.79 -
	Ripoll (Ripollès)	<b>1.34</b> (0.157)
	Sant Hilari Sacalm (La Selva)	<b>2.99</b> (<0.001)
	Blanes (La Selva)	0.50 -
	Santa Coloma de Farners (La Selva)	<b>1.34</b> (0.202)
	Arbúcies (La Selva)	<b>2.46</b> (0.004)
<b>MDS-MPN</b> (N=13) [0.87]	Girona (Gironès)	<b>1.49</b> (0.106)

N=case frequency over the preselected communities; ns=above probability of 0.20 (under Poisson). When the total number of diagnoses is not significant larger than expected (pre-set P-value < 0.05), the series are subject to CUSCORE.

**Table 2 Expected and observed (underlined) number of AML, MDS and MPN diagnoses, Santa Coloma de Farners.**

	AML	MDS	MPN
<b>Reference Years</b>			
1999 E(x)*	0.302	0.484	0.499
2003 E(x)*	0.335	0.551	0.557
2007 E(x)*	0.380	0.635	0.629
<b>Multiyear (1994-2008)</b>			
<b>Expected</b>	4.933	8.074	8.179
<b>Observed</b>	<u>10</u>	<u>14</u>	<u>11</u>

E(x)=expected. \*E(x) of 1999 are used on the sequences for years 1994-1998, and years 1999-2002; thenceforth, E(x) of 2003 and 2007's are used on intervals for years 2003-2006 and 2007-2008, respectively. AML=acute myeloid leukaemia. MDS=myelodysplastic syndromes. MPN=myeloproliferative neoplasms.



**Figure 2. Observed and expected cumulative q-intervals for AML in Santa Coloma de Farners (1994-2008).**

AML=acute myeloid leukaemia. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data indicates increased incidence between the 2<sup>nd</sup> and the 10<sup>th</sup> case.

Cluster for MDS in Santa Coloma de Farners is reflected in Figure 3. Fourteen cases were involved in it, and a permitted 15<sup>th</sup> censored case joined in. Starting from early 1999, the *sq* line (of the observed intervals) had a steeper than expected slope. Inferred *q-interval* values consistently surpassed 0.5, until the final -truncated- interval.

The median age at diagnosis of the patients was 76.5, which compared to the entire province. The by-gender O/E hinted at men bearing a higher excess (Table 10). Mounting up at the dead tail of the cluster, 3 cases of refractory cytopenia with multilineage dysplasia (code ICD-O-3 9985) implied a prevalence of 27.3% amid 11/13 definitively classifiable MDS patients; lightly surpassing the non-clustered towns' contemporary value of 25.7% (Table 11).



5.7). In addition, the aforementioned parallel (time-wise), suspected clustering of MPN patients, deserves mention, if not underpinned for a fact by the exploratory undertakings. The facts, all in all lead to a clear-cut mandate for an in-depth investigation into a causative cluster that overburdened since the 1990s.

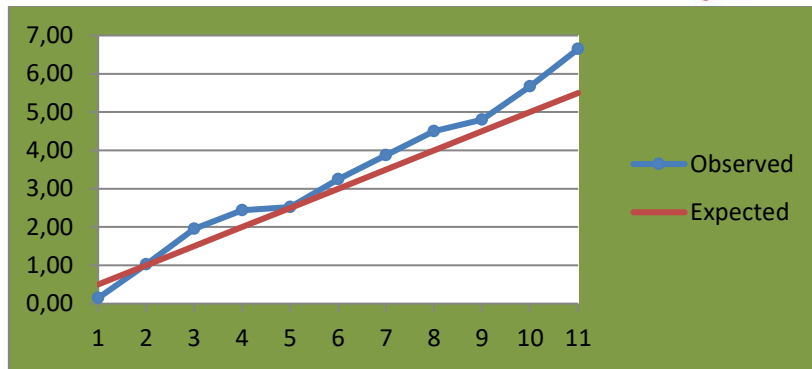
**Table 3 Sequential assessment, CUSCORE. MPN subjects, Santa Coloma de Farners.**

Case Number	Age gender	Date	Interval		
			RI	q	Score
1	76 m	13/11/97	1.9296	0.145	0
2	64 m	17/02/98	0.1303	0.878	1
3	83 f	08/04/98	0.0707	0.932	2
4	66 m	28/09/99	0.7347	0.480	1
5	73 m	10/03/04	2.4644	0.085	0
6	64 m	06/10/04	0.3186	0.727	1
7	69 m	03/08/05	0.4594	0.632	0
8	86 m	07/06/06	0.4702	0.625	0
9	88 m	27/05/08	1.2016	0.301	0
10	38 f	20/08/08	0.1451	0.865	1
11	47 f	01/09/08	0.0192	0.981	2

MPN=myeloproliferative neoplasms. RI (Relative Interval)=Expected number of cases during an observed time interval between cases. q-interval=Null probability that an interval is longer than that observed. Score increases by 1 if  $RI \leq 0.449$  and decreases otherwise, provided score > 0. CUSCORE results are significant at  $p=0.05$ , one-sided if the score reaches 5 at any point of the series.

### 5.1.2 Sant Hilari Sacalm

Another community of The La Selva County bearing temporal clustering resided in Sant Hilari Sacalm, a relatively vast countryside of 83.3 Km<sup>2</sup> and density of 69 inhabitants/Km<sup>2</sup> in 2008. This happened to be a statistically significant MPN cluster ( $E(x) = 4.688$ ;  $P < 0.001$ , Poisson distribution) -Table 1.



**Figure 4. Observed and expected cumulative q-intervals for MPN in Santa Coloma de Farners (1994-2008).**

MPN=myeloproliferative neoplasms. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data indicates fluctuant inconsistent pattern.

Appealingly, the 14 sick individuals included in the analysis, were younger (median age) than all the MPN cases diagnosed in the Province of Girona (59.5 vs. 67). Besides, the disease somewhat deepened among women (Table 10). A conspicuous majority of 12 out of the 14 clustering diagnoses corresponded to individuals with diseases sharing proneness to the somatic JAK2 or allied mutations. Particularly, PV, PMF and ET -ICD-O-3 9950, 9961 and 9962, respectively- amounting to 85.7% of cases. And the comparative prevalence at communities void of clustering was 66.4% (Table 11). During all of the 15 years, the expected figure for these entities had equalled  $E(x)=3.154$ . To bolster such a noticeable gap with the observed count, entertain that the Poisson probability to attain at least 12 patients suffering these JAK2-allied subtypes in Sant Hilari Sacalm during 15 years was 1 in 1000 or smaller. Table 4 shows the raw data set and the standardized intervals. But, missing time data advised a more cautious assessment. The first and third events of the sequence evinced a missing month at diagnosis. In the Table, the calculated range within which the standardized intervals are confined is shown for those with missing month as well as the adjacent, fully dated observations. The expected number of cases for the problematic 6 between 1 January 1994 and 17 November 2004 would be  $E(x)=3.269$ . The probability that the time interval until the 6th case was as short as or shorter than that observed was  $P = 0.113$ , as would be with a cluster on the verge of statistical significance. One wishes, however, not to dismiss best-achievable

information, even if with inaccurate dates. In this context, a legitimate use of cuscore by the strictest conservative approach did the job by talking to all of the concerning data. Hence factoring in longest possible relative intervals upon those missing dates and indeed respecting the null  $RI_{crit} < 0.405$  that pertains to the actual entire data set, the test triggered an alarm on 1 August 2007 (for clearness, procedure is skipped in Table 4). Confirmation held over, until data past 31 December 2008 were achievable.

The  $Sq$  curve describes the temporal cluster (Figure 5). A segment of the plot oscillated in synchrony with that spell surrounding the couple of diagnoses devoid of day and month time, yet followed the trend of the curve. In order to further asses internal validity of this epidemic activity, reckon the following: one would expect  $E(x)=1.874$  MPN patients over the clustering stretch to start -as a proxy- in mid 2003, and continued through the end of the series in 2008. The intensity ( $\gamma$ ) by the rate ratio O/E amounted to 5.3. A supplementary intensity estimate  $\gamma = 7.7$  using the  $q_m$  equation –where

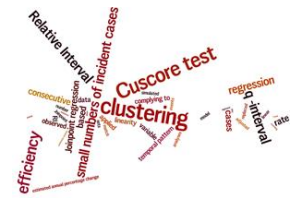
$$\gamma = \frac{\ln(0.5)}{\ln(q_m)},$$

articulated the ballooning rates over 10 consecutive short standardized intervals.

To sum up, a temporal cluster starting by 2003 took its toll of the town community of Sant Hilari Sacalm: an MPN excess of cases with a high risk intensity estimate, pounding up to 7.7 times the baseline rate (even though this magnitude merits a reserve since a 7-fold intensity should be very uncommon if related to an unknown exposure). Arguably, causative exposures as well as acceleration of carcinogenesis may have been expressed via the dropping median age at diagnoses, explicitly, 7.5 net years below the provincial's contemporary figure. The epidemic uncovered as well an above anticipated count of JAK2-allied mutation entities as embraced by the cluster (further evaluation resumed below in [Section 5.7](#)).

### 5.1.3 Arbúcies

Arbúcies, the third municipality of La Selva County, occupies a vast a countryside village (86.2 Km<sup>2</sup>; density in 2008=76 inhabitants/Km<sup>2</sup>). It contained an excess of MM



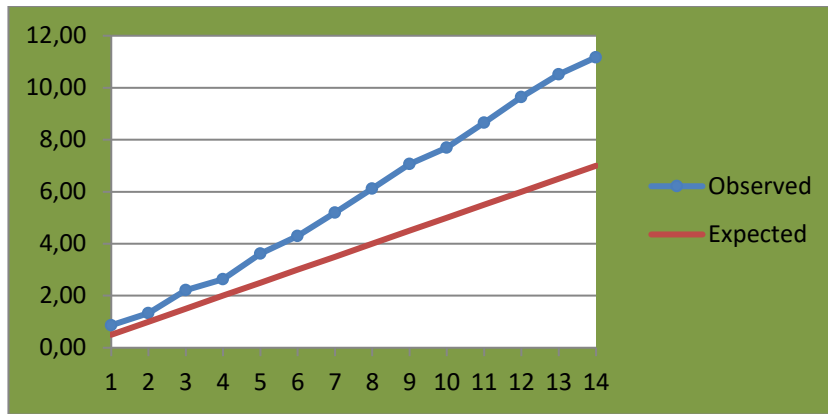
morbidity that eventuated in a statistically significant cluster of MPN ( $E(x)=4.877$ ; Poisson  $P = 0.004$ ) -Table 1.

**Table 4 Sequential assessment. MPN subjects, Sant Hilari de Sacalm.**

Case Number	Age gender	Case date	Interval	
			RI	q
1	49 f	??/??/1994	→	<b>0.0007</b>
			→	<b>0.865</b>
2	63 f	21/02/1997	→	<b>0.2902</b>
			→	<b>0.6214</b>
3	64 m	??/??/2003	→	<b>0.465</b>
			→	<b>0.9108</b>
4	61 f	20/03/2003	→	<b>0.0008</b>
			→	<b>0.887</b>
5	71 m	03/04/2003	→	<b>0.2397</b>
			→	<b>0.0008</b>
6	51 m	17/11/2004	→	<b>0.412</b>
			→	<b>1.7722</b>
7	80 f	08/03/2005	→	<b>0.0008</b>
			→	<b>0.994</b>
8	46 m	08/06/2005	→	<b>0.1116</b>
			→	<b>0.2835</b>
9	42 m	11/08/2005	→	<b>0.668</b>
			→	<b>0.5223</b>
10	80 f	08/03/2005	0.0993	0.905
11	46 m	08/06/2005	0.0805	0.923
12	42 m	11/08/2005	0.0563	0.945
13	58 m	10/01/2007	0.4568	0.633
14	74 m	19/02/2007	0.0376	0.963
15	77 f	14/03/2007	0.0241	0.976
16	56 f	01/08/2007	0.1319	0.876
17	38 f	13/10/2008	0.4160	0.660

MPN=myeloproliferative neoplasms. RI (Relative Interval)= Expected number of cases during an observed time interval between cases; Ricrit=0.405. q-interval=Null probability that an interval is longer than that observed. If a point measure of RI is unachievable, a median q interval is derived out of longest and shortest possible -in bold- This proxy qi values endow a proxy value for graphic approximation.





**Figure 5 Observed and expected cumulative q-intervals for MPN in Sant Hilari Sacalm (1994-2008).**

MPN=myeloproliferative neoplasms. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data indicates increased incidence from the 5<sup>th</sup> subject to the latest.

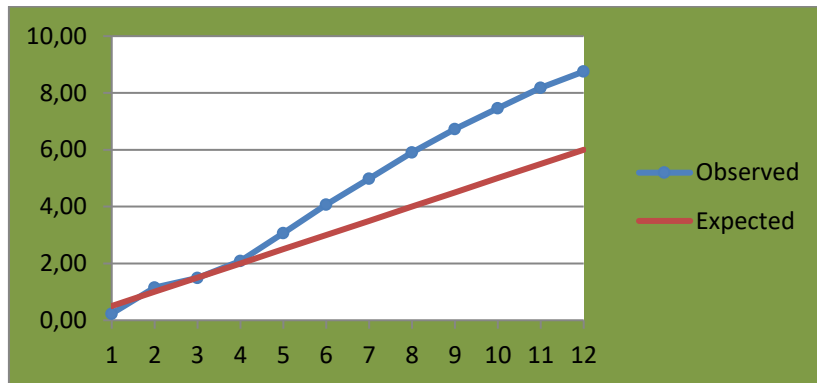
The 12 subjects affected along were older than the total number of patients suffering from MPN in the province (median age: 73 vs. 67). By-gender O/E suggested that male individuals experience the worst (Table 10). Seven out of the 9 clustering patients bore specific entities, namely the JAK2-associated PV, PMF and ET, thereby indicating that they shared common genetic risk markers. Comparing with same subtype counts from non-clustered municipalities, Arbúcies' 77.8% outweighed the 66.4% contemporary prevalence of the former (Table 11). The observable huddle lasted 4.53 years (evaluating from 1 January 2004 to 11 July 2008), and the expected frequency in all those 3 JAK2-associated entities had been  $E(x)=0.992$ . The Poisson probability to observe 7 or more cases in the given time distance was supportive:  $P < 0.0001$ .

The graphical pattern (Figure 6) accounts for the sheer rate rise observed during the latter 5-year period of the sequence; springing out of the 4<sup>th</sup> case.

Concisely, a compelling congregate of MPN patients registered in Arbúcies. This featured a conspicuous 2.46 epidemic intensity, where arguably men fared beyond the usual – hence compatible with a source-like path of causation e.g., overburdening male gender. The median age shifted to the older subjects by 6 years. This supposed procrastination of presentation may indicate distinguishable susceptibilities, which may have since been acquired by the victims<sup>166,177</sup>. The cluster onset took place in early 2004 and continued to the end of the



series. An outbreak of JAK2-associated entities encompassed through its time stretch deserves attention (and I shall resume the issue below in [Section 5.7](#)). Summarily, this countryside findings as well as the neighbour Sant Hilary Sacalm's, pointed to a newly leukaemogenic process obliging in-depth follow up and exhaustive investigation.



**Figure 6 Observed and expected cumulative q-intervals for MPN in Arbúcies (1994-2008).**

MPN=myeloproliferative neoplasms. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data indicates increased incidence from the 4<sup>th</sup> subject to the latest.

#### 5.1.4 Blanes

Blanes, the fourth community with seeming clustering in La Selva, occupies a coastal town of 17.7 Km<sup>2</sup> with relatively high population density (2,214 inhabitants/Km<sup>2</sup> in 2008).

Nineteen individuals were affected with an AML ( $E(x)=15.688$ ) over the 15 years included in this study (Table 1). This yielded an apparent excess of cases above the multiyear expected. Because the O/E fell below statistical significance ( $P > 0.20$ ), the series underwent the waiting time test. The CUSCORE, with a permitted inclusion of a 20<sup>th</sup> censored observation, did not signal (data not shown). The *sq* curve displayed no hint for imbedded clusters either (graph not shown).

#### 5.1.5 Lloret de Mar

Lloret de Mar comprises a coastal town surrounded by the countryside with a total area of 48.7 Km<sup>2</sup> that in 2008 had a density of 775 inhabitants/Km<sup>2</sup>. An O/E rate ratio above one indicated that its community bore a borderline excess of MDS cases (14 observed vs. 13.217 expected; Poisson  $P > 0.20$ ) -Table 1. The CUSCORE did not evince temporal clustering

(scheme not presented); on no account did the observed *sq* curve mirror any huddling during the sequence (not exhibited).

The opening trio of communities were subjected to 4 indicative clusters; a pair of these partially overlapped (time-wise). Besides, a fifth one, and highly suspected MPN clustering at Santa Coloma de Farners deserves a follow up too. Their municipalities of residence, which belong to a common county, are closely located too. The centre of the first municipality, Santa Coloma de Farners, the county's capital, is  $\cong 13$  km (in aerial distance) to the centre of the other 2 towns. These in turn, Sant Hilari de Sacalm and Arbúcies' centres are  $\cong 7$  km away <sup>214</sup>. If one by and by suspects common source carcinogenesis causation, this vicinity had better keep in mind. A geographical look less *prima facie* would lead us to speculate that all associated causal clusters -those described here and in the following- would have taken their toll on incidentally interspersed municipalities. That is, persuading us for the very fact that the communities that supported this epidemic activity were inhabitants of a rather circumscribed area of the south-central part of the province. It will be interesting to explore clustering in some other closely located residential areas.

## 5.2 Gironès County

As seen in Table 1, the Gironès County is second top by number of preselected municipalities with possible clustering for MMs categories over the period of this study. In fact 3 of those did merit sequential analysis herein.

### 5.2.1 Girona town

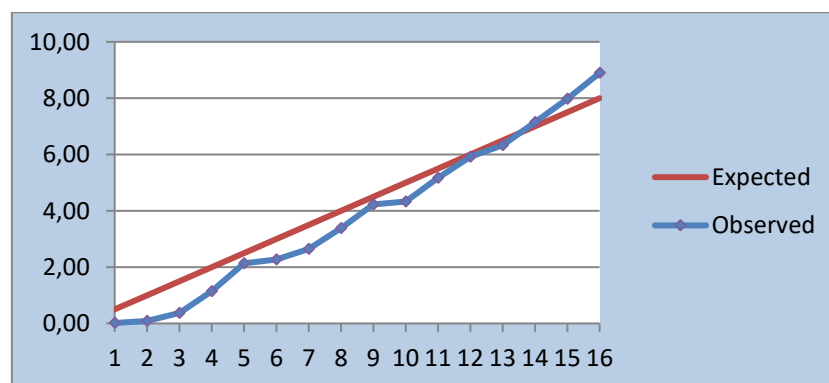
Girona town represents the capital of the county and the province. It reaches 39.1 Km<sup>2</sup> and a population density of 2,415 inhabitants/Km<sup>2</sup> in 2008. In keeping with the O/E figures, the capital of the county seemed to stand up to clustering for each of the four main MM categories – All the attendant probabilities being Poisson  $P > 0.05$  (Table 1).



**Table 5 Sequential assessment, CUSCORE. MDS subjects, Girona town.**

Event	Case Number	Case date	Interval		
			RI	q	Score
1	4	15/04/1996	9.3610	0.016	0
2	8	12/01/1998	7.1230	0.076	0
3	12	10/02/1999	4.4079	0.358	0
4	16	15/09/1999	2.4425	0.770	1
5	20	23/11/1999	0.7725	0.992	2
6	24	24/05/2001	6.1460	0.139	1
7	28	17/06/2002	4.3510	0.368	0
8	32	04/02/2003	2.5788	0.741	1
9	36	07/08/2003	2.0790	0.843	2
10	40	22/03/2005	6.6459	0.102	1
11	44	28/09/2005	2.1130	0.836	2
12	48	10/05/2006	2.5220	0.753	3
13	52	11/05/2007	4.1011	0.414	2
14	56	24/11/2007	2.1926	0.821	3
15	60	03/06/2008	2.1471	0.830	4
16	64	30/10/2008	1.6700	0.911	5(alarm)

MDS=myelodysplastic syndromes. RI (Relative Interval)=Expected number of cases during an observed time interval between events. Event size=4 cases. q-interval= the probability for a longer RI than that observed. Score increases by 1 if  $RI \leq 2.849$  and decreases otherwise, provided score>0. CUSCORE results are significant at  $p=0.05$ , one-sided if the score reaches 5 at any point of the series. Pending confirmatory test.

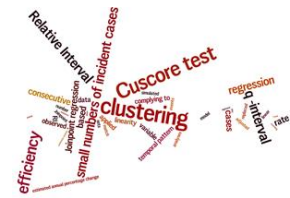


**Figure 7 Observed and expected cumulative q-intervals for MDS in Girona town (1994-2008).**

MDS=myelodysplastic syndromes. The q-interval is calculated for each event (that includes  $r=4$  cases). The curve of the observed data indicates increased incidence at or before the 4<sup>th</sup> even to the latest.

The sequence on MPN cases featured a rate ratio estimate verging on significance (1.21;

$E(x)=62.875$ ;  $P=0.059$ , under Poisson). Yet, an embedded cluster sprang over the last 6 years.

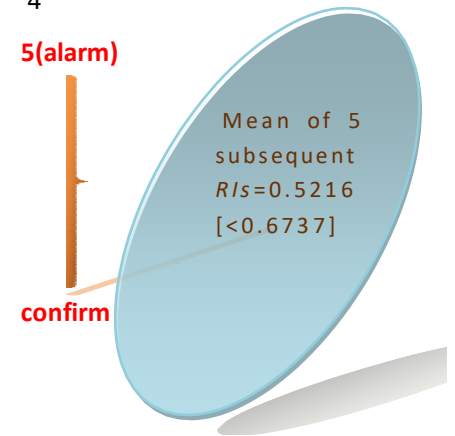


CUSCORE test detected clustering that grew confirmed (Table 6). The table depicts the sequential test scheme and the breakdown of the individual five post-alarm intervals used in the confirmatory test (the shadowed rectangle). The balloon therein contains the result for the mean-based confirmatory test and the reference cut-off point. The sq-plot presented in Figure 8 shows the constant elevated risk since around the 7<sup>th</sup> event. For these 76 patients living in Girona over the 15-year period, the by-gender O/E hinted at a relatively heavier burden on men (Table 10). The median age of those sick people insinuates to fall below their contemporaries in the entire province (61 vs. 67). The revisited  $\gamma$  estimate, taking the epidemic departure from the 7<sup>th</sup> event on, had a value of = 2.02, which represented better the upward departure from baseline rates.

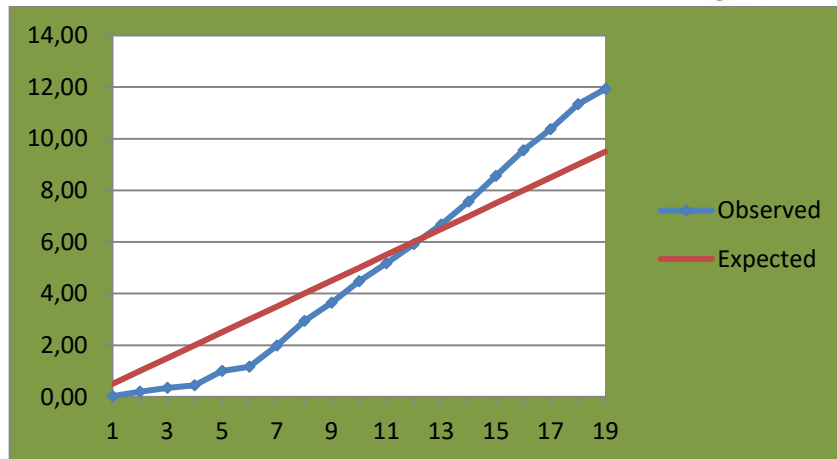
Let us anchor the departure of the clustering at 1 January 2003. Out of 52 clustering cases, 35 (67.3%) had mutated JAK2 associated entities, barely surpassing the prevalence over the clustering-skipped communities in the study period. The other diagnoses included 12 (23.5% vs. 20.9%) Ph+-associated -ICD-O-3 9863 and 9875- (Table 11). As for other morphology subtypes, their so scarce frequencies would not warrant interesting insights either. Given a length of 5.90 years, and  $E(x)=18.013$ , the Poisson probability that no less than 35 or more individuals harboured mutated JAK2 entities was  $P < 0.001$ , which bears the gap present. And the contemporary probability that no less than 12 or more patients contracted Ph+ associated diseases was  $P=0.012$  (or 1 in 83), with  $E(x)=5.596$ .

**Table 6 Sequential assessment, Cuscore and confirmation. MPN subjects, Girona town.**

Event	Case Number	Case date	Interval		
			RI	q	Score
			<b>8.3834</b>		
1	4	??/??/1996		<b>0.027</b>	0
			<b>8.9423</b>		
			<b>5.5424</b>		
2	8	<u>15</u> /06/1997		<b>0.168</b>	0
			<b>6.1012</b>		
3	12	21/11/1998	6.0081	0.150	0
4	16	30/06/2000	6.7416	0.096	0
5	20	19/04/2001	3.3650	0.566	0
6	24	18/09/2002	5.9266	0.158	0
7	28	26/03/2003	2.1890	0.822	1
8	32	23/07/2003	1.3623	0.950	2
9	36	18/03/2004	2.7362	0.706	3
10	40	22/09/2004	2.1424	0.831	4
11	44	20/05/2005	2.7124	0.698	
	45	14/09/2005	1.3274		
	46	20/10/2005	0.4192		
	47	23/11/2005	0.3842		
12	48	30/12/2005	0.4308	0.744	
	49	04/01/2006	0.0466		
13	52	08/08/2006	2.5383	0.749	
14	56	10/01/2007	1.7698	0.896	
15	60	14/02/2007	0.3959	0.999	
16	64	23/04/2007	0.8034	0.991	
17	68	31/10/2007	2.1890	0.822	
18	72	20/02/2008	1.2808	0.959	
19	76	26/11/2008	3.2136	0.599	



MPN=myeloproliferative neoplasms. RI (Relative Interval)=Expected number of cases during an observed time interval between events. Event size=4 cases. q-interval=the probability for a longer RI than that observed; if a point measure of RI is unachievable, a median q-interval is derived out of longest and shortest possible -in bold.- SCORE increases by 1 if RI ≤2.782 and decreases otherwise, provided score>0. CUSCORE results are significant at p=0.05, one-sided if the score reaches 5 at any point of the series. NB the 'nuance' covers a breakdown of the single 5 intervals used in the confirmatory test.



**Figure 8 Observed and expected cumulative q-intervals for MPN in Girona town (1994-2008).**

MPN=myeloproliferative neoplasms. The q-interval is calculated for each event (that include r=4 cases). The curve of the observed data indicates increased incidence that continues from about the 28<sup>th</sup> case.

To sum up, Girona town presented 2 MM-clusters, MDS and MPN. The techniques exposed outsets nearly 1999 and 2033, respectively; and the inflicting intensities were 1.31 and 2.02. The shift to a younger age, among the patients with MPNs -6 years- advises of acceleration and/or exogenous induction of concern. The count distribution of MDS subtypes involved in the Girona town clustering departed from the expected. That is an absolute excess of refractory anaemia with excess of blasts, an entity deemed to embody an overburdening carcinogenicity -highly ‘*pernicious*’ subtype. On the other hand, imperceptible over occurrences of Phi+ and JAK2-related subtypes registered through the MPN huddle. I shall return to deal with this in [Section 5.7](#). One may hypothesize that exposure to whatsoever emergent causative process had been mounting up in this community from around the early to mid-nineties.

### 5.2.2 Cassà de la Selva

Cassà de la Selva is a countryside town (area=45.2 Km<sup>2</sup>; density in 2008=205 inhabitants/Km<sup>2</sup>) that belongs to Gironès County. In keeping with *SIR* results, the community was affected with MDS and MPN clusters (7.655 and 7.272 expected cases respectively), both being significant under the Poisson ( $P < 0.05$ ) -Table 1.

The 14 MDS individuals who were involved in the analysis grew with a greater, if negligible, by-gender O/E among women (Table 10), and just older than their province’s contemporaries were -median age 81 vs. 77. Twelve out of 13 cases engulfed by the cluster have

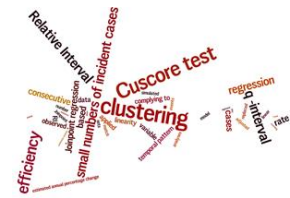


undergone a diagnosis firmly classifiable. Of these, 7 (58.3%) suffering from refractory cytopenia with multilineage dysplasia (code ICD-O-3 9985) joined the cluster – whereas in the clustering-skipped communities, the contemporary figure was 25.7% (Table 11). The corresponding expected value had amounted to  $E(x)=0.916$ ; hence attending a supportive chance in ten thousand or slighter (Poisson), to finding 7 or more of these entities in the clustering period by 11.58 years. Meanwhile, other MDS entities took place in unremarkable prevalence. Figure 9 exhibits the amplified intensity as was denoted by an above 0.5 slope (but for one outlier observation) taking over from the fourth year of the MDS series.

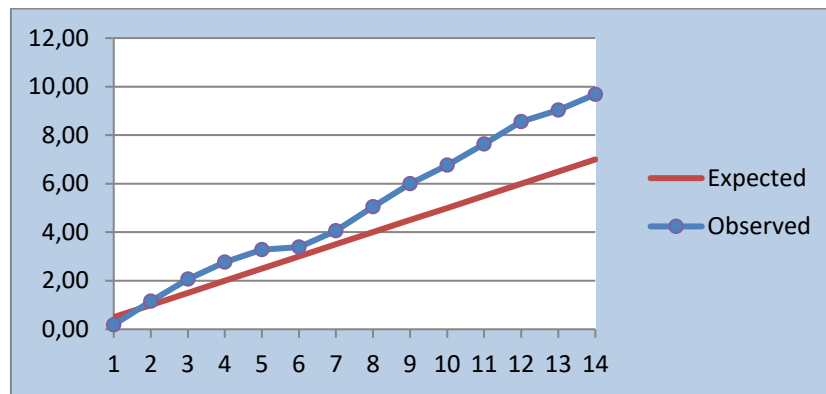
The median age for the 14 MPN patients was higher than the Girona Province (77 vs. 67). Male gender underwent marginally worse O/E ratio. Upon staring at entities accruing by the ending tail of 7 definite clustering cases, there is a prevailing prevalence of the JAK2-associated ones: six patients, that is, an 85.7%, clearly exceeded the contemporary prevalence for the 66.4% sum-up value over the communities devoid of clustering (Table 11). In addition, under the Poisson, and, provided  $E(x)=1.421$ , the probability to observe at least 6 cases through the 4.04 years accruing short intervals was 1 in 100.

As shown in Figure 10, the MPN *sq* curve articulates a temporal cluster that rises with the 8<sup>th</sup> diagnosed case (mid-2004). Because the 6<sup>th</sup> case (year 2003) lacked a precise month at diagnosis, gauging the adjoining intervals elicited unsteady maximum-minimum *RI* values, in the sense that would enact opposite meaning. A ratio of 7 observed cases to the summation of *RI*s from the 8<sup>th</sup> diagnosis on equals 3.37. This bettered the reflection of the intensified rate comparing to the overall rate ratio posited in Table 1.

In short, in the wake of Girona town's community, the people contemporarily living in the nearby Cassà de la Selva (aerial distance  $\cong$  11 km) endured 2 clusters. And these consisted of MDS and MPN, and partially overlapping too. Best-estimated risk intensities were 1.83 and 3.37 respectively. With the MDS sick subjects the increased rates began in 1997, and about 2004 for the MPN patients. The median ages at diagnosis is older than the reference throughout, appreciably among the latter (10 years difference). This may indicate susceptibility contingent



with conditions that worsening with age, for instance, via earlier primary malignancies. Counts of mid-risk MDS subtypes registered exceeding the expected for the cluster length. Likewise, over the observed epidemic sequences of MPNs, the JAK2-associated subtypes did take over (assessment resumed below in [Section 5.7](#)). Summarily, an ongoing and long lasting causative cluster(s) should merit in-depth research in Cassà de la Selva, present since the last decade of the XX century. It can be speculated that a causal cluster of these related leukaemias could have been acting protractedly -with a sustained exposure prevalence- to the point of insulting vulnerable inhabitants.



**Figure 9 Observed and expected cumulative q-intervals for MDS in Cassà de la Selva (1994-2008).**

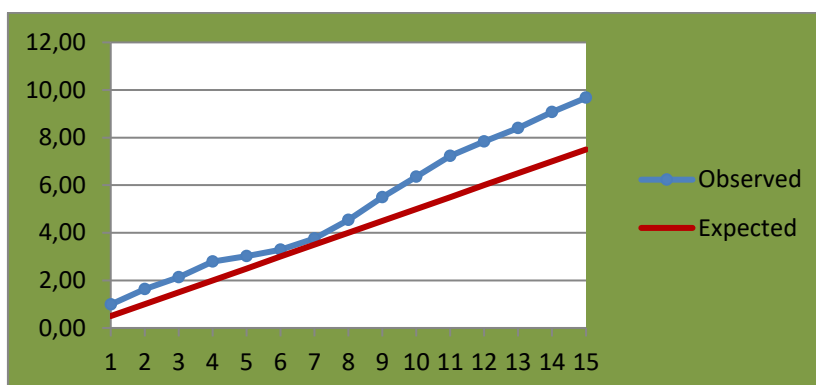
MDS=myelodysplastic syndromes. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence.

### 5.2.3 Salt

Imperceptibly bordering the main city of Girona, Salt stretches for 6.64 Km<sup>2</sup> and was home to 4,333 inhabitants/Km<sup>2</sup> in 2008. Worth noting, despite such proximity, that according to socioeconomic and demographic indicators <sup>204</sup>, Girona will enjoy a conspicuous superior wealth, and population steadiness in terms of inhabitants turnover. This advised that Salt undergo independent analyses, albeit inflated counts.

According to the O/E rate ratio, Salt’s community supported an excess of MDS cases throughout the 15-year period (Table 1). For 19-recorded patients and 17.269 expected, the Poisson probability was clearly not significant ( $P < 0.37$ ) at 0.05 cutting point, hence I continued with waiting time assessment.

The CUSCORE test did not signal for significant cluster. Nor did the *sq* curve expose any imbedded cluster (test scheme and exhibit skipped).



**Figure 10 Observed and expected cumulative q-intervals for MPN in Cassà de la Selva (1994-2008).**

MPN=myeloproliferative neoplasms. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. A 15<sup>th</sup> censored case is included in the plot. The curve of the observed data depicts increased rate from the 8<sup>th</sup> subject.

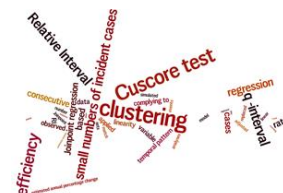
### 5.3 Pla d'Estany County

Turning now to Pla d'Estany County, only the population of its capital, Banyoles, was preselected and further evaluated (Table 1).

#### 5.3.1 Banyoles

Banyoles' territory spreads over 11.05 Km<sup>2</sup>, with a population density of 1,621 inhabitants/Km<sup>2</sup> in 2008. The O/E values of Banyoles' community showed excess of persons diagnosed with either AML or MDS. Both associated Poisson probabilities were > 0.05 (Table 1).

Twelve patients ( $E(x)=8.429$ ) suffering from AML lived in the town. Arguably, the illnesses excess engrossed men, as the by-gender O/E ratio showed (Table 10). Those patients were younger than their contemporaries in the province (median age 61.5 vs. 68.0). The CUSCORE test did not appear to be informative (scheme skipped). The display of the observed *sq* curve described a departure sequence with fluctuating slope followed by an elevation (~2002-2005), if subdued, and a cease. The observed ending-tail encompassed an uninterrupted chain of



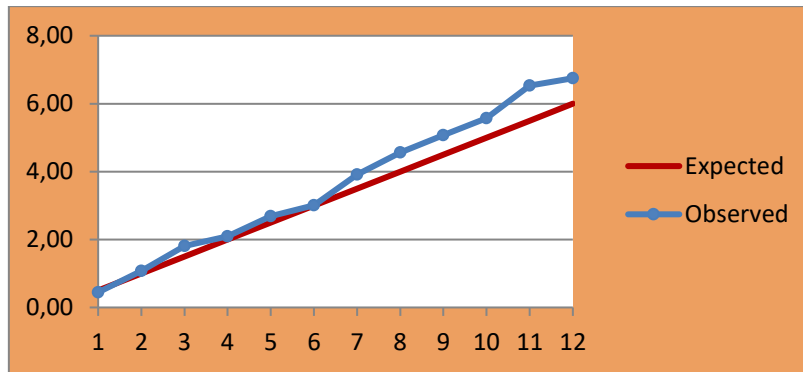
5 sub-mean *RI* intervals (i.e. below *r*) or,  $RI < 1$ , where attending *qi*-values offered above 0.5 figures and a subsequent large interval of 0.22 (Table 7) – hence shouldering a disambiguation of what at a glance gives the sense of a parallel to the expected *sq-plots* (Figure 11).

At that, 4 (80.0%) out of the 5 AML patients subjected to this cluster -assumed ad hoc to last 3.8 years- had diagnoses compatible with proven *de novo* entities, that is to say, a prevalence about level compared to communities void of clustering (83.0%) -Table 11. The pertaining expected figure had been  $E(x)=1.790$ .

**Table 7 Sequential assessment, AML subjects, Banyoles.**

Case Number	Age gender	Case date	Interval	
			RI	q
1	41m	04/07/1995	0.8088	0.445
2	72m	15/05/1996	0.4632	0.629
3	63m	<u>15</u> /12/1996	0.3128	0.731
4	67f	<u>15</u> /04/1999	1.2512	0.286
5	82m	05/04/2000	0.5213	0.594
6	68m	08/05/2002	1.1216	0.326
7	84f	19/07/2002	0.1058	0.900
8	52m	03/05/2003	0.4356	0.647
9	66m	07/07/2004	0.6698	0.512
10	49f	20/09/2005	0.6841	0.505
11	79m	19/10/2005	0.0458	0.955
12	82m	10/05/2008	1.5186	0.219

AML=acute myeloid leukaemia. RI (Relative Interval)=Expected number of cases during an observed time interval between events; RIcrit=0.432. q-interval=the probability for a longer RI than the observed.



**Figure 11 Observed and expected cumulative q-intervals for AML in Banyoles (1994-2008).**

AML=acute myeloid leukaemia. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data depicts rate acceleration imbedded; this expands between the 7<sup>th</sup> and the 11<sup>th</sup> individual.

Eighteen cases ( $E(x)=14.659$ ) inhabitants suffered from MDS in Banyoles. Their age at diagnosis likened to the province's (median age 78.6 vs. 77), and the by-gender O/E denoted a determining incidence on men (Table 10). Sequential analysis of the excess of MDS cases using the CUSCORE, which included a permitted censored case, did not declare clustering (Table 8). Yet a new-found (2004-2008) cluster registered in the *sq-plots*, beginning with the 12<sup>th</sup> case (2004), just after an oscillation of the assessed intervals attributable to one observation with missing month at diagnosis in the same year (Figure 12). The ratio of 7 observed patients by the sum of the corresponding 12<sup>th</sup>-18<sup>th</sup> RI-intervals gave a fair intensity estimate of 1.78

In order to survey into subtype strata, and in fairness to the rule of thumb, I consigned the spawn of the huddling to January 1 2004. Thus, to the last patient of the series with diagnosis dated 8 August 2008, a 4.6-year time-length of elevated rates involved 9 individuals. Four out of the 8 who had a definitely classifiable diagnosis (50.0%), constituted an above current prevalence of refractory anaemia with excess of blasts (ICD-O-3 9983) – being 23.0%. the comparable proportion within non-clustered communities. In addition, there was a chance in 45 for you to observe 4 or more sick persons with these regarded as high risk MDS, given an expected frequency  $E(x)=1.051$  for such stretch of time.

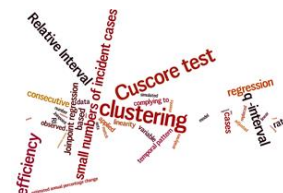
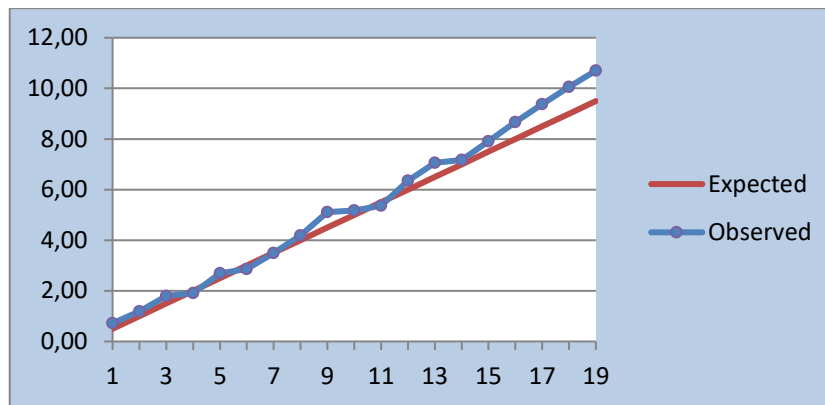


Table 8 Sequential assessment. MDS subjects, Banyoles.

Case Number	Age gender	Case Date	Interval	
			RI	q
1	76 f	11/05/1994	0.3285	0.720
2	83 f	06/03/1995	0.7455	0.475
3	62 f	25/09/1995	0.5029	0.605
4	84 f	20/02/1998	2.1858	0.112
5	76 m	25/05/1998	0.2401	0.787
6	66 f	06/06/2000	1.8472	0.158
7	79 m	07/12/2000	0.4574	0.633
8	81 m	02/05/2001	0.3664	0.693
9	67 m	30/05/2001	0.0708	0.932
10	84 m	??/??/2004	2.4460	<b>0.056</b>
			3.3240	
11	81 m	16/11/2004	0.0025	<b>0.190</b>
			3.3240	
12	78 m	17/11/2004	0,0028	<b>0.997</b>
			0,0028	
13	75 m	14/04/2005	0.2898	<b>0.705</b>
			0.4096	
14	87 f	07/06/2007	2.2050	0.110
15	84 f	16/09/2007	0.3063	0.736
16	92 m	14/12/2007	0.2723	0.762
17	79 m	08/04/2008	0.3528	0.703
18	81 m	11/08/2008	0.3806	0.683
	== CENSORED ==	31/12/2008	0.4332	0.648

MDS=myelodysplastic syndromes. RI (Relative Interval)=Expected number of cases during an observed time interval between cases. Ricrit=0.368. q-interval=Null probability that an interval is longer than that observed. If a point measure of RI is unachievable, a median q interval is derived out of longest and shortest possible -in bold. - These proxy qi values endow an ad hoc value for graphic approximation.



**Figure 12 Observed and expected cumulative q-intervals for MDS in Banyoles (1994-2008).**

MDS=myelodysplastic syndromes. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data depicts rate elevation imbedded; this expands from the 12<sup>th</sup> case.

Arguably, double epidemic has been apparent in Banyoles community: increased rates of AML and MDS. Delineated since around 2002 and 2004 approximately, these 2 distinguishable clusters partially juxtaposed. The pertinent intensity estimates were 1.42 and 1.78, respectively. The AML individuals were 6.5 years younger by the median than the province's contemporary patients had been, and whilst the epidemic spared females, strikingly showed a deeper than expected toll on men, yet again based in small numbers. In unison, the MDS cluster mimicked such trends, but displaying less magnitude. It seems to suggest a profile shouldering acceleration and/or substantial exogenous induction as well as gender-path proneness carcinogenic impairment. Moreover, an attendant observed count excess of *de novo* AML entities fairly duplicated the expected during its cluster's evolution. Simultaneously, the longer MDS epidemic had concurred with a count excess of the highly '*pernicious*'-graded subtype during the huddle's span (further analyses resumed below in [Section 5.7](#)). The -indeed arguable- AML clustering abruptly stopped before the end of the series. Besides, the MDS aggregation of cases, featuring a promineny of the highest '*pernicious*' subtype, persists past and remains at its height towards the end of the study period. Thus, relying on the recognized commonalities between these kindred collection of illnesses, it makes sense that Banyoles' situation was being compatible with a source-like carcinogenic cluster(s) that may had



commenced in the 1990's to proceeding ahead. Such prompt warrants an exhaustive investigation.

## 5.4 Baix Empordà County

Baix Empordà County accounts resulted in a community with excessive cases (Table 1).

### 5.4.1 Palafrugell

Palafrugell is a coastal town with a land surface of 26.9 Km<sup>2</sup> that contains 822 inhabitants/Km<sup>2</sup> in 2008. A non-significant AML rate ratio registered on this community (attendant Poisson  $P = 0.129$ ).

The observed 14 AML cases ( $E(x)=9.903$ ) throughout the 15-year period, were diagnosed at a younger age, the median being 61.5 years, versus the 68 acknowledged for the province. The by-gender O/E was slightly higher among women. The CUSCORE test allowing a censored case does not reveal clustering (Table 9).

Looking at the *sq-plots* in Figure 13, one cannot disregard the existence of 2 main elevated slopes. First, an above expected slope projected from the second case (end of 1995) enduring some 3.5 years; dropped to a nadir of sub-expected *qi* values interspersed among the 8<sup>th</sup>-10<sup>th</sup> patients until past mid-2006, and then it followed an embedded cluster-like on. All of these huddling diagnoses were *de novo* (but not of germline) MMs. Such tail ending of 5 consecutive short intervals (including the censored) did not allow of a conclusion.

Briefly, acute myeloid leukaemias with  $SIR=1.55$  were detected in Palafrugell community. The observed cases throughout the 15-year period, presented their disease at a younger than usual age (median difference=6.5). The increased incidence involved the last 2 years of the series, (encompassing 4 cases and a censored). A throughout upward trend of the *sq-curve* hinted at similar risk elevation – nevertheless, a nadir encompassing 3 successive cases

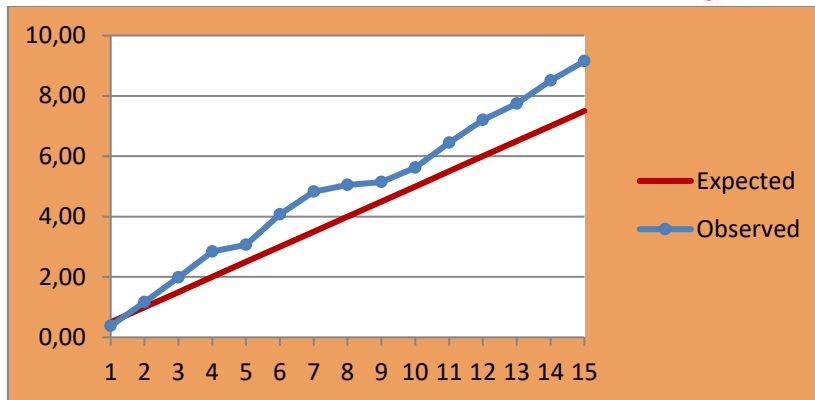


levelled against it. These were patients with accurately informed date of diagnosis. If the AML category of leukaemias is identified by health providers without much procrastination, one might not speculate on missing-time-distance sources – except random thereof. All the same, interpretations must be postponed until incoming incidence data from the cancer registry be accessed, lest the entire epidemic reflection vanished due to small accidental  $qi$  figures.

**Table 9 Sequential assessment, AML subjects, Palafrugell.**

Case Number	Age gender	Date	Interval	
			RI	q
1	40 m	31/07/1995	0.9740	0.378
2	60 m	<u>15</u> /12/1995	0.2307	0.794
3	86 f	20/04/1996	0.2136	0.808
4	76 m	<u>15</u> /07/1996	0.1452	0.865
5	87 f	<u>15</u> /12/1998	1.4866	0.226
6	60 f	<u>16</u> /12/1998	0.0017	0.998
7	37 m	30/05/1999	0.2802	0.756
8	87 f	04/11/2001	1.4935	0.225
9	27 f	06/06/2005	2.3677	0.094
10	51 m	05/07/2006	0.7350	0.480
11	81 m	17/10/2006	0.1927	0.825
12	55 f	08/03/2007	0.2779	0.757
13	78 m	12/01/2008	0.6174	0.539
14	63 f	27/05/2008	0.2742	0.760
	= =CENSORED= =	31/12/2008	0.4346	0.648

AML=acute myeloid leukaemia. RI (Relative Interval)=Expected number of cases during an observed time interval between cases. Ricrit=0.405. q-interval=Null probability that an interval is longer than that observed.



**Figure 13 Observed and expected cumulative q-intervals for acute myeloid leukaemias in Palafrugell (1994-2008).**

The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. An imbedded upturn in rates involves the cases at tail-end.

## 5.5 Ripollès County

Ripollès County has one community with an MM excess of sick individuals. They live in its capital Ripoll (Table 1).

### 5.5.1 Ripoll

The town land area is 73.7 Km<sup>2</sup>. Density=149 inhabitants/Km<sup>2</sup>. A toll of 15 observed MPN individuals registers on, in excess of the expected multiyear ( $E(x)=11.152$ ), being the associated Poisson  $P = 0.157$ .

The possible cluster was not declared by the CUSCORE test (scheme not shown). Nor did the *sq* curve expose any embedded huddle (data not shown).

**Table 10 By-gender O /E by main MM category at clustered communities 1994-2008.**

COMMUNITY	AML		MDS		MPN	
	Male	Female	Male	Female	Male	Female
Santa Coloma de Farners	7/2.89 (2.41)	3/2.03 (1.47)	8/4.46 (1.79)	6/3.61 (1.66)		
Sant Hilari de Sacalm					7/2.53 (2.77)	7/2.16 (3.24)
Arbúcies					9/2.62 (3.44)	3/2.25 (1.33)
Girona town			29/31.27 (0.93)	34/30.08 (1.13)	41/31.25 (1.31)	35/31.62 1.11
Cassà de la Selva.			7/4.20 (1.66)	7/3.46 (2.02)	8/3.86 (2.07)	6/3.41 (1.76)
Banyoles	9/4.78 (1.88)	3/3.65 0.82	11/7.81 (1.41)	7/6.85 (1.02)		
Palafrugell	7/4.69 (1.49)	7/4.20 (1.66)				

AML=acute myeloid leukaemia. MDS=myelodysplastic syndromes. MPN=myeloproliferative neoplasms. The underscored ratios stress the gender with higher excess of cases for a category in a community.

## 5.6 Concurrent independent clusters

The tailored sequential techniques working together have thereby detected 10 indicative temporal aggregations, encircling 6 communities. And by re-counting only those elicited through the tests, 8 indicative clusters showed (5 communities). Of these, 3 harboured coincident and independent MM clusters. And the question arose as to whether or not such twofold malignancy assemblages that had been observed across the local subsets would have been ascribable to plain chance. Of course the question withstands on condition that the inquiry design, if anything, hints no intrinsic bias for these finding entailing the allotment of the 35 series present. One assumes it not be the instance. All the more so because for the Girona province remit, it is apparent that chances are a quantity of communities *not included* in the current study might exhibit clustering. That is, the ones not bearing the pre-set threshold of 10 individuals standing up to a main MM category by 15 years – specifically, should a frequency of affected residents, say 9, outweigh yet the expected and lumps together.



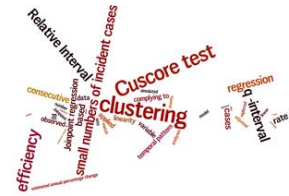
Banyoles, Cassà de la Selva, Girona, Santa Coloma de Farners and Sant Hilari Sacalm. If one meets a consistent departure from the usual occurrences of concerned no-return diseases or between their variants, one will weigh to elucidate the matter further.

Furthermore, an in-cluster subtype (or a group of allied subtypes) predominance could be hypothesized if the clustering-period prevalence out of its pertinent MM category in a clustered community grew greater than the 15 yearly prevalence among <<*non-clustered*>> (or *clustering-skipped*) communities. These encompass all 8 shortlisted towns not yielding any indicative time clustering with the given sequential procedures; as follows: Blanes, Figueres, Lloret de Mar, Olot, Palamos, Ripoll, Sant Feliu de Guíxols and Salt. As a whole, they accounted for 290 MM cases, i.e., half the province's toll. Table 11 summarizes the results.

Odds Ratio (**OR**) and 90% confident intervals were also ran to explore these intriguing associations, which allowed addressing whether there was a consistent pattern for each concerned-ancillary subtype(s) throughout the clustered communities. As can be seen from Table 11, such a pattern arose revolving around 3 biologically meaningful subtypes by the following depiction.

Taking the 4 towns where clustering for MDS took their toll, they did so unfolding predominance of a middle to high-risk subtypes; that is, of the kind ensuing and trailing carcinogenesis disruption at its height. First, Girona's and Banyoles' MDS clustering fared with predominance of the highly '*pernicious*' excess blast subtype. At that, Banyoles' prominent absolute excess of 27.0% deserves mention.

Second, of the in-cluster cases suffering middle-risk (mid- '*pernicious*' ) MDS, namely, the refractory cytopenia with multilineage dysplasia, the period prevalence in Santa Coloma showed a negligible (1.6%) proportion difference to the figure of the non-clustered communities, whereas Cassà de la Selva's surpassed it clearly (32.7%). Overall, an association between clustering status and mid- '*pernicious*' MDS evinced high difference of 17.8% (predominance at the clustered communities) and a strength (OR) of 2.21.



Third, 4 communities endured indicative temporal clusters of MPN. All of which entangled concerned-ancillary subtypes in increased proportions. Among them, Girona solely bore Ph+ associated diseases overrepresentation. This was a period prevalence though inconspicuously (proportion difference=1.2%) exceeding the figure for the non-clustered communities. On the other hand, Girona town inhabitants together with other 3 communities were crippled with predominance for JAK2-associated subtypes. The aggregated data thereof equalled an absolute excess of 6.8% in prevalence and an association strength (OR)=1.38 in favour of an in-cluster predominance, given such a subtypes.

**Table 11 Period prevalence by malignancy subtypes among the shortlisted communities with and without indicative clustering, Girona Province, 1994-2008.**

Subtypes		<u>Clustered Communities</u>		<u>Non-Clustered</u>		Odds Ratio
		n/N	Prevalence{%} [90% CI]	Prevalence{%} (N)[90% CI]		
<b>De novo AML</b>	Banyoles	4/5	80.0 [34.7-88.3]	83.0 (88) [75.4-88.5]		
	Sta. Coloma	8/9	88.9 [62.3-97.5]			
<b>MDS-EB</b>	Banyoles	4/8	50.0 [24.9-75.1]	23.0 (74) [16.0-31.9]	1.28	
	Girona	9/39	23.1 [14.9-35.7]			
	Total MDS-EB	13/47	27.7 [18.4-39.4]			
<b>MDS-MLD</b>	Cassà de la Selva	7/12	58.3 [35.6-78.0]	25.7 (74) [18.3-34.8]	2.21	
	Sta. Coloma	3/11	27.3 [11.5-52.0]			
	Total MDS-MLD	10/23	43.5 [28.1-60.3]			
<b>MPN- JACK2</b>	Arbúcies	7/9	77.8 [50.4-92.4]	66.4 (128) [59.3-72.9]	1.38	
	Cassà de la Selva	6/7	85.7 [54.8-96.7]			
	Girona	35/52	67.3 [56.0-76.9]			
	Sant Hilari	12/14	85.7 [64.7-95.2]			
	Total MPN-Jack2	60/82	73.2 [64.5-80.4]			
<b>MPN-Ph+</b>	Girona	12/52	23.1 [14.9-33.9]	21.9 (128) [16.5-28.4]		

n/N=count of subentities per myeloid malignancy category; MDS-EB=Refractory anaemia with excess of blasts; MDS-MLD=Refractory cytopenia with multilineage dysplasia; MPN-JACK2=Janus Kinase 2 gene mutation subtypes; MPN-Ph+= Philadelphia chromosome-associated myeloproliferative neoplasms. The period for the non-clustered communities lasts 1994-2008, and the observed epidemic activity span for those clustered.



## 6. DISCUSSION

Research and Public health intervention in the realm of preventable no-return diseases suppose the sensible recognition and then the preparedness to tackle their clustering. It is implied that even global public health problems may well underlie small or ‘one-off’ nonrecurring causal clusters <sup>215,216</sup>.

Features such as inherent rarity, group place and time confines, or otherwise methodological mandates such as diagnosis ascertainment, confirmation and stratification, impose aggregations into low counts of events. Besides the low frequency attribute proper, other uncertainties rise upon the investigation of such huddles: missing dates of disease presentation, fluctuant registration completeness or medical practice over the years, and timeliness of data (e.g., when achievability hinges on periodical census of the reference population or when high turnover impacts the ascertainment by residency); at that, the exhaustion of susceptible persons -in a community with low turnover rate- that baffles the results, among others <sup>21,63,109</sup>.

Several ‘post-analytical’ challenges supervene, which one had better plan ahead of time. For example, the ever-present possibility of chance to the attribution of clustering discovery; unsettled sizes of the sample's universe to trace the sample error of the study <sup>36,37</sup>. Or else, the invisible multiplicities sources of comparisons per sample unit; even worst, the Texas sharpshooter fallacy – contesting a researcher due to his tidying up data on causes and consequences after-the-fact <sup>48,217</sup>.

With several of such issues, non-ex-post pre-specified investigational frameworks withstand better than those motivated by alleged cluster or *ex-post* undertakings. Yet by all odds the former would not truly preclude them. After all, true randomness may not fulfil the construct of disease causation <sup>3,217</sup>. This is whence responsible and responsive authorities are supposed to equip themselves. No sooner officers lack reassuring methods, than they draw hesitant



conclusions. You chance to find a partaking of such parlance on certain published guides about huddle investigations <sup>21</sup>.

Apropos of the purpose of not dithering or postponing let us assert that detection of temporal clusters of no-return rare diseases may be delayed upon just measuring general secular trend statistics. In this dissertation, tailored waiting time analyses were integrated for being implemented on high quality morbidity data from a population-based cancer registry. It yielded indication for arguably true clusters and clued in next causative epidemiological queries. Such implementation in pre-hoc given series of infrequent cancers entitled novelty while exposing far-reachingness as an exploratory approach.

A published report on time trends in myeloid leukaemias from a whole province confronted me <sup>113</sup>. It covered modelled standardized incidence rates for the 4 comprehensive categories of these malignancies. Derived estimated annual percentage changes showed incidence rates that remained steady for acute myeloid leukaemia and significantly growing for the other 3 MM categories <sup>113</sup>. Therefore the report, descriptive as it stood, warranted public health scrutiny. Namely, that causative clues be steadfastly elucidated – by answering to what time and space distribution those rare diseases had happened.

Noteworthy, the crude data set produced by the Girona province's population-based cancer registry, qualified for fitness-for-use (the registry bear IARC's standards) <sup>210-213</sup>. So much so that you were able to marshal sequential analyses to covering any community that recorded at least 10 observed sick persons with any of those 4 main MM categories during 1994-2008. Across categories, yield case-occurrence ranged 0.87-1.33 per community-year. This is persuasive enough to support the necessity for sequential procedures based on non-fixed time intervals, sensitive to events interspersed and throughout deployed; namely, at ease with small numbers settings so that significance tests will not have an ultimate a bearing on the derivation of conclusions. Definitely, providing that the time of departure from sub-epidemic stages requires scrutiny.



Throughout the RESULTS section, I proceeded with interpretations of the compelling clustering as per community and sometimes over their locations. Aptly, these walkthroughs entitle to put forward which community risk issues require decisions; which most likely point to a causal relationship of the cases, or whether a public health problem might be ruled out. By the same token, I explored the results of the analyses executed respecting the in-cluster prominencies for concerned cancer subtypes. Such a deepening be worthwhile provided that it lead to workable causal hypotheses. That is, to the extent that it be revealed that during epidemic activity, patients whose cancer subtypes bear an epidemiological meaningful connotation prevail. Here, the posited direction of inquiry works sidelong to those recently published by Du et al. (2019). These authors instead explored associations of opiate adverse effects and interesting predictors by comparing severely affected patients entangled in clusters to those spared from spatial-temporal aggregations. The proportion of patient's characteristics (such as a cancer condition) with a putative bearing on such clusters was so enquired <sup>218</sup>.

Quizzing person variables relies on a breakdown of this paucity of observations, which inevitably elicits small numbers for analysis. Consequently, results interpretability thereof cannot progress void of admonition. It stands reasonable, however the downside, to maximize a workable inspection of biomedical available data. Just because this is on all accounts what a public health agent is summoned to upon these front investigations: cut delays, and refrain from ignoring nor the wood neither the trees. Suffice is to say that interlinking the thus far softer, 'doubtful' suspects and the more solid outputs stemming from analyses on larger observations often eases and deepens the readability and scopes of each other. Biological plausibility of mechanistic viewpoints aids in the reasoning; and this is whence visiting the library is always worthwhile <sup>3,219–221</sup>.

The community of Banyoles provided with a demonstration case as follows. This intermediate size community showed, in the onset, 2 'no significant' MM clusters with relatively low intensity rise. By inspecting the graphic temporal pattern, as well as the integration of supplementary results, albeit built from scarce observations, it led to a tenable



grasp. Just as strengthening the case for green light towards a more exhaustive clustering investigation and hinting also in favour of a source-like carcinogenic cluster hypothesis.

As far as decision-making is concerned, responsible authorities will frequently operate under the judgment of morbidity elevation. Yet again, many decisions ought to be supported, denied or earnestly held until more information becomes available. One ought to have insisted that this warrants the consideration of several aspects in concert. In between the decision-making process, public health agents had better contemplate some clues regarding the possible responsible causative instance.

Thus for example, temporal clustering combined with spatial clustering (as reflected by clustering in some bordering towns) may indicate exposure to a close factory emitting a carcinogenic agent. Similarly, excessively large O/E in one gender had better lead to propose a source of causation. Certainly, breakdown of these data by gender would hardly effects clues. On the other hand, a striking variation under epidemic signs deserves attention, and on no accounts left disregarded as though an incident or outlier. The community of Arbúcies presented a compelling cluster of MPN; here, the discrepancy between men's O/E of 9/2.62 and the women's 3/2.25 is exemplificative, thereby should be audited towards next decisional steps.

As regards the MM subtypes, the exploratory flavour of the appraisals present, justified extracting all underlying content to their variations, taking advantage of their well-confirmed morphological diagnoses. Of which, one might well have tried any breakdown; but it would be genuine if the polytomous classes of the subtypes, as well as the questions posed, be established from the outset. This, however, is not easy, firstly because of the nature or the design-imposed small counts; secondly, and for the time being, due to the paucity of scientific-empirical background serving the actual realm. In the end, I sought whether pre-selected concerned subtypes expressed beyond custom throughout the clustering span, and univariately analysed differences in period prevalence between communities subjected and not, to observable risk

elevations (i.e., experiencing or not epidemic activity with any MM whatsoever). Workable epidemiological insights grew enhanced, somehow, from such fine-grained examination.

Now I am to address putative paybacks stemming from the ascertainment of *specificity of effect* to the benefit of the actual exploratory phase. Nevertheless it should be kept in mind that not very many pathognomonic endpoints can be recalled in occupational and environmental cancer epidemiology <sup>163</sup>. Ecologically investigated, incidence is often elevated throughout several cancer types. Examples may include yet familial-driven huddles, as suggested at a large population-based study <sup>222</sup>. For the advancement of this knowledge, Carroll et al. (2017) assessed the added value of *shared* covariates among both frequent and much less frequent respiratory malignancies in mixture models of risk estimations <sup>223</sup>.

By now, one has been anticipating that carcinogenic effects are likely to be reflected on clustering of diagnoses beyond the ‘regular noise.’ Nevertheless, the reason why a carcinogenic activity might lead to express a subtype(s) propensity at a supervening agglomeration of cases is unknown.

On behalf of public health operability, I selected concerned-ancillary subtypes (or grouped kindred subtypes or entities). Briefly, the selection contemplated: First, whether the entities (or allied ones) matched a *taxonomically firm* subtype, for which there was also epidemiological bedrock supporting exogenous aetiologies, for instance the Phi+ associated subtypes. Second, it contemplated whether the entity (ditto) coalesced into an authoritative construct for a gradient of any detriments in carcinogenic development – in fact, uniquely featured among the MDS entities (see the revised 4th edition of WHO’s Classification of LHs)

116.

With respect to the aforementioned preselected subtypes and by the time confines of the definite clusters herein, I should highlight the following intriguing finding: a detection of absolute prevalence excesses, and/or Poisson-significant incidence outweighing the expected, concurred for most of the ratified clusters. Yet, many of forthcoming questions will not help but render ourselves devoid of ascertainable answers. For instance, are those consistent



overexpressions pronouncing alike causation paths? Else; might an endured evenness of subtypes (i.e., relative frequencies' distribution spared as usual) as well being expressing a rather heterogeneous hazardous cocktail, or, complex mechanistic paths outlining host-environment interactions; both – as current 'regular noise'? In my view, the chance to detect the contribution of one subtype to an encountered cluster is quite slim, first by dint of the small numbers, and second because of individual host factors may increase susceptibility to one or another subtype. At present, an indication to association with one subtype may be obtained from several towns. Granted one would not expect more than a vague clue, this approach is expounded below.

The 4 communities that have undergone indicative MPN clustering as well depicted that the increased intensities were largely contributed by either Ph+ or mutated JAK2-associated cases on top. Regarding clusters featuring the latter, the excess, if of variable magnitude, deserves further attention. These huddles appeared simultaneously during the last half a decade of the study and seemed to maintain intensities beyond 2008. One of those communities experienced an above-expected case count harbouring Ph+, contained as well by the same epidemic stretch. In this context, it is worth reminding that longitudinal studies have shed light into the evidence of environmental aetiology of MPN subtypes associated with *JAK2*<sup>33</sup>. In addition, there are no germline related predispositions agreed for the Phi+ associated subtypes; on the other hand, acute radiation exposure, has been causatively implicated<sup>116</sup>.

Of the MDS indicative clusters detected, a post hoc excess of cases of refractory cytopenia with multilineage dysplasia was identified in the long-lasting temporal clusters at Santa Coloma de Farners and Cassà de la Selva. These happened to feature rather moderate intensities. Besides, the unusual by-gender O/E insinuated (apiece) source-like of causative cluster. Abiding by WHO's endorsed nomenclature, that midway-detriment (or 'pernicious')-graded entity of MDS is deemed to be driven by lifelong leukaemogenic buildup<sup>116,130</sup>.

Refractory cytopenia with multilineage dysplasia, for the record, has been attributed to exogenous carcinogens <sup>224</sup>. Interestingly, being capable to pinpoint the 2008' WHO-morphological subtypes of LHs, convincing risk excess with benzene exposure raised at an occupational case control study on a sample of hundreds of cases <sup>225</sup>. Namely, an exposure-response pattern for all MDS and that subtype in particular. At the end of the day, the finding of an excess risk sentinelled by such class of entities in a long-term exposure occupational environment is corroborating the widespread truism that chronic exposure to leukaemogenic factors, acting stochastically via somatic mutations, brings about leukaemias.

Myelodysplastic cases with excess blast clustered and juxtaposed past the spell of an embedded likely time clustering of de novo AML in the town of Banyoles. The adding up of the compelling commonalities between AML's and that deemed highly-*'pernicious'* MDS entity, to their mere duplicated agglomeration may well boost the case for the verisimilitude of an incoming causation cluster for MMs in this community.

Furthermore, I have detected an overlap of embedded-like agglomeration of MDS individuals with excess blast and JAK2-associated MPN ones. Both indicative huddles render the mandate to investigate causation cluster: arguably mechanistically coalescing to harm the inhabitants of the capital of the province.

For the sake of progress in public health prevention, so far, one renders puzzled by the real utilitarian extent of specificity and fine-tuning that revolves around MMs. Specific classes actually present intra-construct limitations throughout. A clear-cut illustration is the so-called 'secondary' AMLs following progression from an MPN, a primary MDS or primary MDS-MPN. It may be unfeasible to distinguish these natural disease evolutions from iatrogenic-induced changes, which in fact are merged as a unique subtype, namely, 'Therapy-related myeloid neoplasms' <sup>116,167,226</sup>. Thereby actual incidence turns out to be an underestimate due to misclassification.

After all one might not overemphasize what the leitmotif of experts coming up with these taxonomies has been. Let us retail the WHO reference; if enunciating the usefulness of its



categorizations it clearly states: <<. . .in daily practice for therapeutic decision-making [and the provision for] a flexible framework for integration of new data.>>[sic] <sup>116</sup> – more than would in the elucidation of exogenic commonalities, or further comprehending subtypes by their apportioning at epidemic circumstances.

So much for my interpretation attempts on the results yielded from omics-based refinements. To the best of my knowledge, albeit, a scarcity of empirical data features the domain of morphological entities regarding MMs – either aimed to prompt a cluster investigation or elucidate an ascending risk thereof. In any case, the subjection of cancer epidemiology, just as for referral agencies to such levels of specificity of effect had long since rendered subsidiary. Moreover, causation paradigms endorsing an otherwise ‘non-specificity’ are taking over by the growing acceptance of multiple mechanistic pathways. Then, noxious agents and causative instances are being inquired or ascribed to definitely broad endpoints; among others, ‘anaemia’ -where aplastic anaemia mixes- ‘(upper) aerodigestive tract’, ‘non-Hodgkin lymphoma’ ‘Leukaemia’, ‘Leukaemia, childhood’ or at most ‘leukaemia, acute myeloid’. Nor are researchers refraining from recalling broadly defined malignancies upon recollecting the pros of so called ‘precision’ medicine <sup>158</sup>.

## STRENGTHS AND WEAKNESSES

The authors of the report that prompted this dissertation had put forward an upward significant secular trend in 3 out of the 4 MM categories. In their letter to the editor <sup>113</sup>, notwithstanding, they challenged the reliability of these increased incidences –as if had been caused by advances in diagnosis. This dissertation’s results do not second this, despite its identical data setting. Suffice is to say that had an enhanced-diagnosis bias been effecting the calculated rates, they would have manifested everywhere in the covered region; undoubtedly, in the populations closest to the haematological referral centres – through which underestimated



rates would have generated series of spurious short intervals between cases. At any rate, this has not been the case.

Secular changes in the type of reporting provider (outpatient vs. hospital), or in the biomedical accuracy and diagnostic procedures, presumably lead to under or overreporting <sup>227</sup>. Logically, the more the characterization of a disease occurs outside the hospital walls, the more likely the failure in the registry coverage. This obliges case-finding endeavours by the registry centre. But had expected rates biased downwardly, systematic errors leading to the impression of clustering would have emerged. In this dissertation, the absence of mimicked outbreaks or patterns across the different samples discarded suspicion of systematic mistakes of substance revolving around the referenced rates.

An immigration process determining demographic impacts started in Catalonia in the mid-90's and has typically comprised young people of working age. Consequently, up to 2008, one estimates a negligible influx, in terms of aging, into the group over 64 years old (the age stratum MMs diagnoses were registered throughout the period) <sup>204,228</sup>. This should have otherwise been of concern, because of the relatively short latency periods of the LHs –whose secular disease rate changes would have mirrored bias by migration.

In general, residential towns are quite stable (in size and demographic profile) over 15-years periods. Some of the scrutinized, however, experienced considerable turnover. Factoring in overall immigration and emigration over 2005-2006, a 25% or more occupants of Salt, Lloret de Mar, Girona and Blanes managed to migrate out or into the municipality. On account of misclassification (of place at diagnosis), it potentially veers from the likelihood of clustering detection as hedged by these administrative localities. For reasons of latency periods, and the aforementioned skewness in the age distribution of the actual MM-series, the migratory dynamics have though a far-fetched impact.

Conceivably, too, one should warrant an admonition in estimating the timing of occurrence for malignancies and perhaps other non-return chronic maladies. Expectedly, should an elevation of epidemic intensity show up, it would start moderately. This is rooted in the



pattern. This is whence one must refer to the natural history of these entities. Certainly, in keeping with haematologists' practices and resources contemporary to the actual series, the MPN illnesses lent themselves more to either underascertainment or unsteady diagnosis dates than other LHs do <sup>5,152</sup>. Here also is framed an illustration on how uncertainty of disease registration modifies the cumulative score responsiveness and (to a lesser degree), the strength of the graphical depiction.

To return to the advantages of the complementarity between these procedures, the *sq* curve may lend support to the assumption that the alarm is true before data for confirmation is available. An example to support of clustering as yet before confirmation was seen in the MDS cluster that signalled for Girona Town. In the figure the curve indicates that at least from the 8<sup>th</sup> event (except for one event), the slope of the curve is larger than 0.5. Upon an instance like this, the chance arises for the responsible authorities to start pondering clues sooner than a median-based confirmation test become feasible.

On the whole the *sq* curve's possible looks have been outlined here and elsewhere <sup>48,107,217</sup>. By virtue of the data scenarios in focus, zigzagging patterns are still a probable appeal. How should one persevere and audit the point in change in  $\lambda$  regardless? In doing so, you wish you lean on epidemiologically sound reasoning. An approximation is to concede a stress to the length of the observed run of intervals displaying a consistent direction of the slope -above or below the expected. Let us consider the remit of the settings present, where the yearly-expected number of diagnoses is  $< 1$  (or  $r=1$ ) and the significance test, too inflexible, has not declared clustering. This thereby summons the accumulative *qi* curve for a revisit. And insofar as the aim becomes not to miss an otherwise genuine clustering, it commissions us to ameliorate any disturbances of an elevated slope. At present, a delay on waiting time that does not overly exceed an *RI* unit means an interval lag close to the anticipated run for one case to arise. It seems a reasonable a priori guide that a downward turn of the curve stemming from a gap not as long as  $1.5 \times RI$  units be reckoned as an artifact. Otherwise, merits a rewind as if an inflection point in trend. As to *q-interval* terms, a measurable variability emanates from a delay of such a magnitude. For instance, an  $RI = 1.0$ , is  $0.3679 qi$ ;  $RI = 1.5$  gives  $0.2231 qi$ . The former falls

100



between the interquartile range and the first quartile whereas the latter already falls under the first quartile limit of the  $qi$  (uniform) population distribution. On the other hand, affirming that a slope line denotes increased risk, would withstand criticism, as long as it persists for a sizeable spell. Presumably enough, a true cancer cluster would not disappear from an open community after a handful of years.

The graph exhibited for AML in Palafrugell series is of utility: a single observation accounted for the first ‘drop’ among a span that likened to an alarming period (i.e., adjacent short  $RIs$  so that the cumulative score accrues). It took place through a case distanced by  $1.49 \times RI$  units, then, according to the proposed rule, should not uphold an actual nadir or breakup of the increased rate. The next interruption downward instead, took 3 successive long waiting time observations, the second of which lengthened  $2.37 \times RI$  units, thereby readable as a nadir proper. In sensing this way the underlying information of the plot, the quantity of concerning slopes of epidemic relevance is simplified. Lastly, the ending stretch showing an elevated slope registered during last 2 years of the study. That is, too short a spell to declare confidently an epidemic deploy or an advent of such. Alternatively, a tantalizing approach might emphasize a probabilistic reading of, for instance, the likelihood of finding such and such number of stretches with elevated slopes (would factor in 3 instead of 2 of such on those of Palafrugell). Or else in the same way, claiming the dissuasive effect of a little, a priory probability to find a dominant number of adjacent short  $RIs$  throughout the whole study period; and thus reasoning that mere chance or diagnostic deviation had a bearing on the performance of the CUSCORE.

Undoubtedly, the  $sq$  curve still stands as only as useful as its underlying construct allows. And insofar as one keeps in mind a priori intelligence, while heuristically pursuing a "talking to the data", it lends itself for outright interpretability.

Returning to uncertainty about the current procedures, naturally they are responsive to inbuilt randomness as need be too –insofar as one cannot reassure that one is not observing without error<sup>48,107,217</sup>.

Through a beforehand-length-set of 15-years, these techniques working together have thereby detected 10 indicative temporal aggregations affecting 6 municipalities. Using tests *only*, 8 indicative clusters showed otherwise. Of course, municipalities with less population that lagged upon preselection still could have been engulfed by clustering. For example, should 9 cases have observed in a community in which 5 were to be expected. On the other hand, the rate of significant results could have been minor than that observed if selection by the number of cases were avoided.

By now, taking the null- attendant proportion at an universe of 15 communities across 4 main MM categories, the expected number of false alarms is 3; truly  $15 \times 4 \times 0.05$ . Granted, the sample universe was framed with reference to either of possible series that met a cut-off point of  $\geq 10$  count for any of the 4 categories. In all fairness, 35 be the number of such possible series attainable from the reference population. Thus, a chance account for expected significant test of clustering should be given by  $(35 \text{ series}) \times (0.05) = 1.75$  expected clusters, which is less than the observed. All in all one can state that at least some of the alarms are likely to be correct regardless of the chance.

A *positive* confirmatory test lessens the yet present chance of spurious clustering. That is, provided achievability to incoming incidence information is warranted too. Precisely, it weeds out three fourth of the superfluous significant results. After all, functioning on incoming new cases, remains an at hand aid to such a prior objective. Not least, it can fulfil so by a tenable holding time.

Lastly, the problem of multiple comparisons inherently spares here in large degree. This is consigned by the pervading fact that analyses are based on single samples; as though 1 community equals 1 sample unit. Moreover, if one had approached an ex post analysis, one would have found a less simple equation than that mentioned in the foregoing paragraph –



*clarity and transparency of the method, its statistical power to detect the cluster of interest, and the method's ability to produce the desired output.>>*

The approaches so far presented in this dissertation do come across in compliance with the aforementioned tenets. One is comfortable to state that these straightforward but robust methods lend their detection outputs, corollaries and clues to be adopted by public health agents and the people.

## OTHER RESEARCH -REAL DATA SETS

The present research realms should have achieved a proper space in medical research. It failed to fulfil so. Suffice is to attempt the query [*space-time clustering*] at PubMed®, to automatically retrieve hundreds of irrelevant results engendered from a handful of nowhere medical subject headings terms. Most of these comprising either modalities of multivariate methods or interdependence models ('hierarchical clustering' 'cluster analysis') or sampling procedures ('cluster sampling'). Such output is not surprising since to date, The National Library of Medicine search interface is rooted into unfortunate hierarchical trees thereof. Moreover, its search interface yields 380 (rounded) citations for the terms [temporal clusters or clustering] which constitutes 24 and 14 times less than the elicited for queries using either [Voltaren® or diclofenac], or [Codeine], respectively, after excluding the adverse effects of each of these deleterious but lasting products (Boolean search January 1 1977 June 21 2020; [Appendix 3](#)). Furthermore, temporal clusters -or clustering- peer-reviewed articles are tapering year after year <sup>230</sup>.

As regards the reply to the 'why bother' question, plainly several authors have emphasized the far-reachingness of cluster investigations once the results summon such endeavours. This entitles health and scientific payoff encouraging further undertakings to proceed and scour for aetiology. It has been upheld that even if no single factor grows responsible for the rate elevation, the overall attempt frames tipping points for intervention and prevention. Tom Rivers (New Jersey) childhood cancer cluster had proved for 1979-1995; yet





was no reasonable way to state that the results had indicated genuine elevated rate that caused 7 cases in a single year whereas just 1 case registered in each of the 2 preceding and the succeeding 4 years (R. Chen, personal and appointed scientific communication, Universitat Politècnica de Catalunya, April 27, 2012).

A Danish study aimed at identifying space-time clustering for specified types of Non-Hodgkin Lymphoma using population-based registry data (1999-2003). By then, high incidence trends at the national and global levels prompted their research. Granted the focus was on cases' topology, the work lends itself well to contrast approaches. Briefly, procedures included a case-control-cancer cluster statistics that takes into account changing addresses by period; and, for confirmatory assessment, the authors used spatial scan software '*SpaceStat*'. The statistic '*Q-statistics*' relies on stringent threshold  $P$  value (0.001) to signal time aggregation as well as a lenient (0.05) for simultaneous space-time clustering test. This being computed every time a case changed household, ensuing after the evaluation of the hypothesis of no clustering through Monte-Carlo techniques, whereby a range of possible  $P$  values result. <sup>198</sup>

Firstly, one wonders to what extent epidemiological intelligence remains aligned with such complex methodological efforts. Notice how in that research the time span of the study population was limited to 5 years – a never presumptive catchment period to detect cancer clusters. Again, refraining from the information reflected on graphical exhibits displays frustrates the commitment of spotting temporary clusters that could prove genuine. Secondly, the researchers imposed *ad hoc* cut-off-values and bore procedures without spelling them out. Hence, a reader may render either unnoticed of the judgement of underlying assumptions, or mystify by the ramifications of some of these impositions on the interpretation of the results. For instance, the shape of the spatial collection of cluster -in the paper, the authors mentioned an elliptical shape. However, one notes the avoidance of any prior explanation for the rationale underlying such arrangement of scanning windows. And thereby you ask whether another scanning configuration would better the seizing of true clusters <sup>233</sup>. Users set as well the magnitude of the nearest neighbour number of cases to declare clustering. Whenever authors confront such metric devoid of informative hints, they must appeal to repeating tests in order to



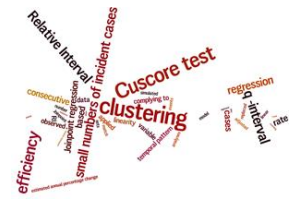
explore the assignable levels of it. Thus, the probabilities from the analyses necessitate corrections for multiple testings<sup>234</sup>. Abiding by authors' discussion, a main drawback resided in the intense computational load due the large data set of participants' topology<sup>198</sup>.

Subjecting the performance of the integrative approach used in this dissertation to comparison trials is not straightforward. Further than the one-off cluster determinants that undermine repeatability, myriad of issues preclude equitable comparisons. So much so that upon contrasting monitoring control charts exclusively, researchers pursue identical data scenarios, or, if comparing akin methods, pre-fixed parameters too -among others: the expected number of patients that are observed until a signal ensues; the odds of the outcome, or the skewness of the patient's risk distribution<sup>235-237</sup>. Yet it is worth noting, that such contrapositions have been referred to monitoring arrangements (in the just cited assessments, risk-adjusted derivatives of CUSUM or THE SETS); which results in relatively more predictable or controllable than those of the tests present.

As a test, readiness and likely effectiveness of CUSCORE stay abreast. And so does it for monitoring applications. Independent methodological studies comparing performance between competing surveillance methods and under small baseline probabilities of event, have likened CUSCORE to due alternatives (e.g. CUSUM-derived control charts)<sup>95,98</sup>. Bear in mind that applicability extends beyond out-of-control rates of disease as such. In retrospect, strong incentives to accomplish potent surveillance methods have occasionally proceeded from failures in medical practice that echoed through an entire health care system<sup>238,239</sup>. Given a tolerable rate of false positives, the optimum be an alarm elicited at minimal delay.

Lastly, research leading to the development of temporal cluster studies on rare, chronic diseases and small homelands should be vigorously fostered<sup>215</sup>. There is a case call for optimizing the exploitation of routine health statistics collected using population-based, public-funded and well-consolidated health registries. Definitely jointly sparked with information enhancement by electronic linkage of existing sources<sup>240</sup>. Nevertheless, central administrations

must enforce that real-time-on site data on exogenous hazards be accessible. By and by, the most critical determinant of validity for environmental epidemiological studies still remains the assessment of exposure <sup>3,14,21,48,184</sup>. Focused interest-geographic area surveillance interwoven with clustering explorations has been postulated as a cost-aware approach. An added payoff for health agencies be a worth of field data for resourcefully interacting with worried communities and communicate with the public <sup>18,215</sup>.



## 7. CONCLUSIONS

1. In the setting of exploratory, non-post-hoc investigations, the CUSCORE test coupled with visual inspection of the temporal pattern of the events disambiguate outputs of incidence of low-frequency chronic diseases suspected to be at an elevated risk.
2. This assemblage of the CUSCORE test and the graphical display, supply the detection and interpretation of time clusters, had these imperceptibly been lying embedded in a period of several years.
3. Because of little incidence and the short period, no data were available for confirmatory analyses in almost all communities. In view of the prospects of false alarms, the confirmatory procedure is indeed necessary.
4. The *accumulative q-interval* enables us to ‘talk to data,’ and that endows us with interpretative cleverness at the same time as compliance with epidemiological reasoning. It affects marginally the number of false alarm. Furthermore, the curve primarily increases the chance to detect true alarms when the test fails.
5. Upon unearthing the *when-about*s intensified rate commenced, these sequential procedures ease a timely scrutiny respecting the intensity degree, fine-grained subtypes, and other hints to the benefit of robustness of the insights
6. The sensitivity of the procedures has been demonstrated even for expected sub-unity event counts a year in communities with few thousands residents, just as for mild-intensities raises beyond non epidemic or baseline reference.
7. Whereas the selective assessments using these rather costless methods have focused on cancer huddles, this approach can be extended to other no-return maladies.
8. Insights achieved as plausible causation understandings of the signalled cogent clusters, prefigure what ought to be investigated next, and outreach to mending interventions and pre-empting chronic maladies.

## 8. REFERENCES

1. Editorial team. Disease clustering: hide or seek? *Lancet* 336, 717–718 (1990).
2. Simpson, B. W., Truant, P. & Resnick, B. A. Stop and listen to the people: an enhanced approach to cancer cluster investigations. *Am. J. Public Health* 104, 1204–1208 (2014).
3. Rothman, K. J. A Sobering Start For The Cluster Busters' Conference. *Am. J. Epidemiol.* 132, (1990).
4. Kolstad, H., Lynge, E., Olsen, J. & Sabroe, S. Occupational causes of some rare cancers: a literature review. *Scand. J. Soc. Med. Supplement*, 1–148 (1992).
5. Seaman, V. *et al.* A multidisciplinary investigation of a polycythemia vera cancer cluster of unknown origin. *Int. J. Environ. Res. Public Health* 7, 1139–1153 (2010).
6. Juzych, N. S. *et al.* Adequacy of state capacity to address noncommunicable disease clusters in the era of environmental public health tracking. *Am. J. Public Health* 97 Suppl 1, 163–169 (2007).
7. Fernández-Navarro, P., García-Pérez, J., Ramis, R., Boldo, E. & López-Abente, G. Industrial pollution and cancer in Spain: an important public health issue. *Environ. Res.* 159, 555–563 (2017).
8. Goodman, M. *et al.* Cancer cluster investigations: review of the past and proposals for the future. *Int. J. Environ. Res. Public Health* 11, 1479–1499 (2014).
9. Massachusetts Department of Public Health; U.S. Centers for Disease Control and Prevention; The Massachusetts Health Research Institute. *The Woburn Environment and Birth Study Synopsis*. (Massachusetts Department of Public Health, 1998).
10. Coglianò, V. J. *et al.* Preventable exposures associated with human cancers. *J. Natl. Cancer Inst.* 103, 1827–1839 (2011).
11. Abrams, B. (Vivi) *et al.* *Investigating suspected cancer clusters and responding to community concerns: Guidelines from CDC and the Council of State and Territorial Epidemiologists. Morbidity and mortality weekly report-Recommendations and Reports* 62(RR08), (Centers for Disease Control and Prevention, 2013).
12. Maslia, M. L. *et al.* Public health partnerships addressing childhood cancer investigations: case study of Toms River, Dover Township, New Jersey, USA. *Int. J. Hyg. Environ. Health* 208, 45–54 (2005).
13. Kristensen, P., Hilt, B., Svendsen, K. & Grimsrud, T. K. Incidence of lymphohaematopoietic cancer at a university laboratory: a cluster investigation. *Eur. J. Epidemiol.* 23, 11–15 (2008).
14. Iva Hertz-Picciotto. Environmental Epidemiology. in *Modern Epidemiology* (eds. Rothman, K. J., Greenland, S. & Lash, T. L.) 598–619 (Lippincott Williams, 2008).
15. Richter, E. D. & Laster, R. The Precautionary Principle, epidemiology and the ethics of delay. *Int. J. Occup. Med. Environ. Health* 17, 9–16 (2004).
16. Stiglitz, J. E. Freedom to Choose? in *Globalization and Its Discontents* 53–88 (W. W. Norton & Company, Inc., 2002).
17. Moriconi Bezerra, M. (I)Legalidad y desmaterialización de la justicia: Consideraciones preliminares sobre su efecto en la estabilidad psíquica del ciudadano. *Rev. Mex. Análisis Político y Adm. Pública* 4, 9–28 (2015).
18. Wartenberg, D. Investigating disease clusters : why, when and how? *J. R. Stat.*



- Soc. . Ser. A ( Stat. Soc. )* 164, 13–22 (2001).
19. U.S. Marine Corps Base Camp Lejeune Disaster. Excerpt from a public meeting by ATSDR, Veterans Administration and Community Assistance Panel. (U.S. Department of Health and Human Service/Agency for Toxic Substances and Disease Registry [URL:<https://www.dropbox.com/s/rqd6l391o3e8uov/MeetingMay122005.swf?dl=0>], 2015).
  20. Agency for Toxic Substances and Disease Registry. *Safeguarding Communities from Chemical Exposures*. (Department Of Health & Human Services. USA; American Public Health Association, 2009).
  21. Saunders, P., Kibble, A. & Burls, A. Investigating clusters. in *Oxford handbook of public health practice* (eds. Guest, C., Ricciardi, W., Kawachi, I. & Lang, I.) 148–157 (Oxford University Press, 2013).
  22. Fleming, L. E., Ducatman, A. M. & Shalat, S. L. Disease clusters: a central and ongoing role in occupational health. *J.Occup.Med.* 33, 818–825 (1991).
  23. Harvey, F. *et al.* *Independent Oversight and Advisory Committee for the WHO Health Emergencies Programme Interim report on WHO 's response to COVID-19*. (WHO, 2020).
  24. Forman, R., Atun, R., McKee, M. & Mossialos, E. 12 Lessons learned from the management of the coronavirus pandemic (in press). *Health Policy* xxx–xxx (2020). doi:10.1016/j.healthpol.2020.05.008
  25. Blouin-genest, G., Bogic, A., Blouin-genest, G. & Champagne, E. *WHO Global Response to COVID-19. Communicating Risk/Risky Communication, Rapid Results Report. Phase I: December 21, 2019 to January 31 2020*. (2020). doi:10.20381/hkzz-an46
  26. Antequera Baignet, J. *Propuesta Metodológica Para El Análisis De La Sostenibilidad Regional*. (Universitat Politècnica de Catalunya. Institut de Sostenibilitat, 2012).
  27. Comisión para Reducir las Desigualdades Sociales en Salud en España. *Avanzando hacia la equidad: propuesta de políticas e intervenciones para reducir las desigualdades sociales en salud en España*. Ministerio de Sanidad y Política Social (Ministerio de Sanidad, Servicios Sociales e Igualdad. Secretaria Técnica, 2015).
  28. Cuzick, J. Clustering. in *Encyclopedia of Biostatistics* (eds. Armitage, P. & Colton, T.) 2, 942–951 (John Wiley & Sons, 2005).
  29. Dong, Y., Hedayat, A. S. & Sinha, B. K. Surveillance strategies for detecting changepoint in incidence rate based on exponentially weighted moving average methods. *J. Am. Stat. Assoc.* 103, 843–853 (2008).
  30. Rothman, K. J. Clustering of disease. *Am. J. Public Health* 77, 13–5 (1987).
  31. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, 209–220 (1967).
  32. Ruckart, P. Z., Bove, F. J. & Maslia, M. Evaluation of exposure to contaminated drinking water and specific birth defects and childhood cancers at Marine Corps Base Camp Lejeune, North Carolina: A case–control study. *Environ. Heal.* 12, 104 (2013).
  33. Gross-Davis, C. A. *et al.* The role of genotypes that modify the toxicity of chemical mutagens in the risk for myeloproliferative neoplasms. *Int. J. Environ. Res. Public Health* 12, 2465–2485 (2015).
  34. Kreis, C. *et al.* Space-Time clustering of childhood leukemia: evidence of an association with ETV6-RUNX1 (TEL-AML1) fusion. *PLoS One* 12, 1–15

- (2017).
35. Besag, J. & Newel, J. The detection of clusters in rare diseases. *J. R. Stat. Soc. Ser. A (Statistics Soc.* 154, 143–155 (1991).
  36. Thun, M. J. & Sinks, T. Understanding cancer clusters. *CA. Cancer J. Clin.* 54, 273–80 (2004).
  37. Schulte, P. A., Ehrenberg, R. L. & Singal, M. Investigation of occupational cancer clusters: theory and practice. *Am. J. Public Health* 77, 52–56 (1987).
  38. Draper, G. J. & Parkin, D. M. Cancer Incidence Data for Children. in *Geographical and Environmental Epidemiology: Methods for Small Area Studies* (eds. Elliot, P., Cuzick, J., English, D. & Stern, R.) 63–71 (Bookcraft L.t.d. Midsomer Norton. Avon, 1996).
  39. Rothenberg, R. B., Steinberg, K. K. & Thacker, S. B. The public health importance of clusters: a note from the Centers for Disease Control. *Am. J. Epidemiol.* 132, S3-5 (1990).
  40. Chen, R. Exploratory analysis as a sequel to suspected increased rate of cancer in a small residential or workplace community. *Stat. Med.* 15, 807–816 (1996).
  41. Chen, R., Connelly, R. R. & Mantel, N. Analysing post-alarm data in a monitoring system in order to accept or reject the alarm. *Stat. Med.* 12, 1807–1812 (1993).
  42. Chen, R. The cumulative q interval curve as a starting point in disease cluster investigation. *Stat. Med.* 18, 3299–3307 (1999).
  43. Neutra, R., Swan, S. & Mack, T. Clusters galore : insights about environmental clusters from probability theory. *Sci. Total Environ.* 127, 187–200 (1992).
  44. Greenland, S. Applications of Stratified Analysis Methods. in *Modern Epidemiology* (eds. Rothman, K. J., Greenland, S. & T, L.) 283–302 (Lippincott Williams & Wilkins, 2008).
  45. Ottino-Loffler, B., Scott, J. G. & Strogatz, S. H. Evolutionary dynamics of incubation periods. *Elife* 6, 1–28 (2017).
  46. Armenian, H. & Lilienfeld, A. The distribution of incubation periods of neoplastic disease. *Int. J. Epidemiol.* 99, 92–100 (1974).
  47. Lawson, A. B. Commentary: assessment of chance should be central in investigation of cancer clusters. *Int. J. Epidemiol.* 42, 448–449 (2013).
  48. Coory, M. D. & Jordan, S. Assessment of chance should be removed from protocols for investigating cancer clusters. *Int. J. Epidemiol.* 42, 440–447 (2013).
  49. Assunção, R. Statistical assessment of cancer cluster evidence-in search of a middle ground. *Int. J. Epidemiol.* 42, 453–5 (2013).
  50. McElvenny, D. M. *et al.* Investigating and analysing workplace clusters of disease: a health & safety executive perspective. *Occup. Med. (Chic. Ill).* 53, 201–208 (2003).
  51. Boffetta, P. *et al.* False-positive results in cancer epidemiology: a plea for epistemological modesty. *J. Natl. Cancer Inst.* 100, 988–95 (2008).
  52. Maslanyj, M., Lightfoot, T., Schüz, J., Sienkiewicz, Z. & McKinlay, A. A precautionary public health protection strategy for the possible risk of childhood leukaemia from exposure to power frequency magnetic fields. *BMC Public Health* 10, 673 (2010).
  53. Editorial team. *Twelve Late Lessons. Late Lessons From Early Warnings : The Precautionary Principle 1896 – 2000. Environmental Issue Report* (Luxemburg Office for Official Publications of the European Communities, 2001). doi:10.1136/oem.59.11.789-a
  54. Miller, A. B. *et al.* Risks to health and well-being from radio-frequency radiation emitted by cell phones and other wireless devices. *Front. Public Heal.* 7, 1–10 (2019).



55. Goldstein, B. D. The Precautionary Principle and Scientific Research Are Not Antithetical. *Environ. Health Perspect.* 107, A 594-A 595 (1999).
56. Gardner, M. J. *et al.* Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. *BMJ* 300, 423–429 (1990).
57. Gardner, M. J. Paternal occupations of children with leukemia. *Br. Med. J.* 305, 715 (1992).
58. Laurier, D. & Bard, D. Epidemiologic studies of leukemia among persons under 25 years of age living near nuclear sites. *Epidemiol. Rev.* 21, 188–206 (1999).
59. Bross, I. D. J. The Primacy Principle [letter]. *Environ. Health Perspect.* 21, 329–331 (1977).
60. Merrifield, M. & Kovalchuk, O. Epigenetics in radiation biology: a new research frontier. *Front. Genet.* 4, 1–16 (2013).
61. Baulch, J. E., Aypar, U., Waters, K. M., Yang, A. J. & Morgan, W. F. Genetic and epigenetic changes in chromosomally stable and unstable progeny of irradiated cells. *PLoS One* 9, 1–13 (2014).
62. Lampe, N., Breton, V., Sarramia, D., Sime-Ngando, T. & Biron, D. G. Understanding low radiation background biology through controlled evolution experiments. *Evol. Appl.* 10, 658–666 (2017).
63. Chen, R. The Efficiency of the CUSCORE Test as Compared to that Applied to SMR in Detection of a Carcinogenic Exposure. in *Asbestos: Risks, Environment and Impact* (eds. Soto, A. & Salazar, G.) 2012, 153–160 (Nova Science Publishers, Inc, 2009).
64. Lambertini, M. *et al.* Cancer and fertility preservation: international recommendations from an expert meeting. *BMC Med.* 14, 1–16 (2016).
65. McHale, C. M. *et al.* Assessing health risks from multiple environmental stressors: moving from  $G \times E$  to  $I \times E$ . *Mutat. Res. - Rev. Mutat. Res.* 775, 11–20 (2018).
66. Boshuizen, H. C. & Greenland, S. Average age at first occurrence as an alternative occurrence parameter in epidemiology. *Int. J. Epidemiol.* 26, 867–872 (1997).
67. Greenland, S. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *Am. J. Public Health* 89, 1166–1169 (1999).
68. National Research Council. *Committee to Review Studies of Possible Toxic Effects From Past Environmental Contamination at Fort Detrick.* (National Academies Press, 2012).
69. Bhatia, S. & Sklar, C. Second cancers in survivors of childhood cancer. *Nat. Rev. Cancer* 2, 124 (2002).
70. Kumagai, A., Reiners, C., Drozd, V. & Yamashita, S. Childhood thyroid cancers and radioactive iodine therapy: necessity of precautionary radiation health risk management. *Endocr. J.* 54, 839–847 (2007).
71. Brown, A. P. *et al.* The risk of second primary malignancies up to three decades after the treatment of differentiated thyroid cancer. *J. Clin. Endocrinol. Metab.* 93, 504–515 (2008).
72. Shilkrut, M., Belkacemi, Y. & Kuten, A. Secondary malignancies in survivors of breast cancer: how to overcome the risk. *Crit. Rev. Oncol. Hematol.* 84, 86–89 (2012).
73. Martin, M. G., Welch, J. S. & Walter, M. J. Therapy related acute myeloid leukemia in breast cancer survivors, a population-based study. *Breast Cancer*



- Res. Treat.* 118, 593–598 (2012).
74. Ståhl, O. *et al.* Risk of birth abnormalities in the offspring of men with a history of cancer: a cohort study using Danish and Swedish national registries. *J. Natl. Cancer Inst.* 103, 398–406 (2011).
  75. Bove, F. J., Ruckart, P. Z., Maslia, M. & Larson, T. C. Evaluation of mortality among marines and navy personnel exposed to contaminated drinking water at USMC base Camp Lejeune: a retrospective cohort study. *Environ. Heal.* 13, 10 (2014).
  76. Azim, H. A. *et al.* Safety of pregnancy following breast cancer diagnosis: a meta-analysis of 14 studies. *Eur. J. Cancer* 47, 74–83 (2011).
  77. Jørgensen, K. T. *et al.* Socio-demographic factors, reproductive history and risk of osteoarthritis in a cohort of 4.6 million Danish women and men. *Osteoarthr. Cartil.* 19, 1176–1182 (2011).
  78. Tough, S., Tofflemire, K., Benzies, K., Fraser-Lee, N. & Newburn-Cook, C. Factors influencing childbearing decisions and knowledge of perinatal risks among Canadian men and women. *Matern. Child Health J.* 11, 189–198 (2007).
  79. Rothman, K. J., Greenland, S. & Lash, T. L. *Modern Epidemiology*. (Lippincott Williams & Wilkins, 2008).
  80. Thoreau, H. D. *The Journal of Henry David Thoreau. II The Journal of Henry David Thoreau* 1, (Dover, 1962).
  81. Lilienfeld, D. E. ‘The greening of epidemiology’: Sanitary physicians and the London epidemiological society (1830-1870). *Bull. Hist. Med.* 52, 503–528 (1978).
  82. Inskip, H. Standardization Methods. in *Wiley StatsRef: Statistics Reference Online* (eds. Balakrishnan, N. *et al.*) (Wiley Online Library, 2014). doi:10.1002/9781118445112.stat06116
  83. Ocaña-Riola, R. Common errors in disease mapping. *Geospat. Health* 4, 139–154 (2010).
  84. Spiegelhalter, D., Grigg, O., Kinsman, R. & Treasure, T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *Int. J. Qual. Health Care* 15, 7–13 (2003).
  85. Shields, P. G. & Harris, C. C. Molecular epidemiology and the genetics of environmental cancer. *J. Am. Med. Assoc.* 266, 681–687 (1991).
  86. Chen, R. & Fromm, P. The CUSCORE test and the q-interval in cluster analyses of colon cancer and of lymphoma among asbestos workers. *Stat. Med.* 22, 3101–3109 (2003).
  87. Ugarte, M. D., Ibáñez, B. & Militino, A. F. Modelling risks in disease mapping. *Stat. Methods Med. Res.* 15, 21–35 (2006).
  88. Richardson, S., Thomson, A., Best, N. & Elliott, P. Interpreting posterior relative risk estimates in disease-mapping studies. *Environ. Health Perspect.* 112, 1016–1025 (2004).
  89. Cançado, A. L. F., da-Silva, C. Q. & da Silva, M. F. A spatial scan statistic for zero-inflated Poisson process. *Environ. Ecol. Stat.* 21, 627–650 (2014).
  90. Sherlaw-Johnson, C. *et al.* Continuous monitoring of emergency admissions of older care home residents to hospital. *Age Ageing* 45, 71–77 (2016).
  91. Steiner, S. H., Cook, R. J., Farewell, V. T. & Treasure, T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 1, 441–452 (2000).
  92. Wolter, C. Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen’s sets technique. *Methods Inf. Med.* 26, 215–219 (1987).
  93. Lawrance, R. A. *et al.* Use of cumulative mortality data in patients with acute



- myocardial infarction for early detection of variation in clinical practice: observational study. *BMJ* 323, 324–327 (2001).
94. Chen, R. A surveillance system for congenital malformations. *J. Am. Stat. Assoc.* 73, 323–327 (1978).
  95. Sego, L. H., Woodall, W. H. & Reynolds, M. R. A comparison of surveillance methods for small incidence rates. *Stat. Med.* 27, 1225–1247 (2008).
  96. Chen, R., McDowall, M., Terzian, E. & Weatherall, J. *Eurocat Guide to Monitoring Methods for Malformation on Registries, EEC Concerted Action Project Eurocat*. (European Economic Community, 1983).
  97. Chen, R. The relative efficiency of the Sets and the Cusum techniques in monitoring the occurrence of a rare event. *Stat. Med.* 6, 517–525 (1987).
  98. Ramaraj, S. & Subramanian, B. D. Surveillance methods for the rare health events—a systematic review. *Int. J. Stat. Distrib. Appl.* 2, 76–80 (2016).
  99. Munford, A. G. A control chart based on cumulative scores. *J. R. Stat. Soc. Ser. C (Applied Stat.)* 29, 252–8 (1980).
  100. Chen, R. & Goldbourt, U. Analysis of data associated with seemingly temporal clustering of a rare disease. *Methods Inf. Med.* 37, 26–31 (1998).
  101. Kordysh, E., Bolotin, A., Barchana, M. & Chen, R. Cancer Epidemiology of Small Communities : Using a Novel Approach to Detecting Clusters (Conference Paper). in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (eds. Crespo, J., Maojo, V. & Martin, F.) 2199, 126–132 (Springer Verlag, 2001).
  102. Cox, D. R. & Lewis, P. A. W. The Statistical Analysis of Series of Events. in *Methuen's Monographs on Applied Probability and Statistics*. 285 (Methuen and Co Ltd, 1966).
  103. Cox, D. R. & Hinkley, D. V. *Theoretical Statistics*. (Chapman and Hall, 1974). doi:10.1007/978-1-4899-2887-0
  104. Lake, A. M. & Gould, M. S. Suicide Clusters and Suicide Contagion. in *A Concise Guide to Understanding Suicide: Epidemiology, Pathophysiology, and Prevention* (eds. Koslow, S. H., Ruiz, P. & Nemeroff, C. B.) 52–61 (Cambridge University Press, 2014). doi:https://doi.org/10.1017/CBO9781139519502
  105. Cheng, Q., Li, H., Silenzio, V. & Caine, E. D. Suicide contagion: a systematic review of definitions and research utility. *PLoS One* 9, (2014).
  106. CDC. *Guidelines for Investigating Clusters of Health Events-Appendix. Summary of Methods for Statistically Assessing Clusters of Health Events. Morbidity and mortality weekly report-Recommendations and Reports* (Department Of Health & Human Services. USA; American Public Health Association, 1990).
  107. Chen, R. & Bitchatchi, E. Y. Detection and estimation of the increasing trend of cancer incidence in relatively small populations. *Cancer Epidemiol.* 50, 207–213 (2017).
  108. Chen, R., Mantel, N., Connelly, R. R. & Isacson, P. A monitoring system for chronic diseases. *Methods Inf. Med.* 21, 86–90 (1982).
  109. Chen, R., Iscovich, J. & Goldbourt, U. Clustering of leukaemia cases in a city in Israel. *Stat. Med.* 16, 1873–1887 (1997).
  110. Anderson, A. *Business Statistics for Dummies*. (John Wiley and Sons Inc., 2013).
  111. Griffon, N. *et al.* Searching for rare diseases in PubMed: a blind comparison of Orphanet expert query and query based on terminological knowledge. *BMC Med. Inform. Decis. Mak.* 16, 101 (2016).
  112. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* 28, 165–173

- (2020).
113. Osca-Gelis, G. *et al.* Population-based incidence of myeloid malignancies: fifteen years of epidemiological data in the province of Girona, Spain. *Haematologica* 98, (2013).
  114. HAEMACARE Working Group. Manual for coding and reporting haematological malignancies. *Tumori* 96, 1–38 (2010).
  115. International Agency for Research on Cancer. & World Health Organization. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. (International Agency for Research on Cancer, 2008).
  116. Swerdlow, S. H. *et al.* *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. *World Health Organization Classification of Tumours* (International Agency for Research on Cancer, 2017). doi:10.1017/CBO9781107415324.004
  117. Vianna, N. J. *et al.* Hodgkin's disease: Cases with features of a community outbreak. *Ann. Intern. Med.* 77, 169 (1972).
  118. Alexander, F. E. Clustering and Hodgkin's disease. *Br. J. Cancer* 62, 708–711 (1990).
  119. Grufferman, S. Clustering and aggregation of exposures in Hodgkin's disease. *Cancer* 39, 1829–1833 (1977).
  120. Russi, M. B. *et al.* Cancer. in *Textbook of Clinical Occupational and Environmental Medicine* (eds. Russi, M. B. *et al.*) 727–824 (Elsevier, 2005). doi:10.1016/B978-0-7216-8974-6.50034-8
  121. Liu, B. *et al.* Myelodysplastic syndromes spatial clusters in disease etiology and outcome. *Leuk. Lymphoma* 1–8 (2015). doi:10.3109/10428194.2015.1071487
  122. Besag, J. & Newel, J. The detection of clusters in rare diseases. *J. R. Stat. Soc. Ser. A (Statistics Soc.* 154, 143–155 (1991).
  123. Boddu, P. C. & Zeidan, A. M. Myeloid disorders after autoimmune disease. *Best Pract. Res. Clin. Haematol.* 32, 74–88 (2019).
  124. Oh, Y.-J. *et al.* Mutation of ten-eleven translocation-2 is associated with increased risk of autoimmune disease in patients with myelodysplastic syndrome. *Korean J. Intern. Med.* 35, 457–464 (2019).
  125. Gupta, R., Webb-Myers, R., Flanagan, S. & Buckland, M. E. Isocitrate dehydrogenase mutations in diffuse gliomas: Clinical and aetiological implications. *J. Clin. Pathol.* 64, 835–844 (2011).
  126. Medeiros, B. C. *et al.* Isocitrate dehydrogenase mutations in myeloid malignancies. *Leukemia* 31, 272–281 (2017).
  127. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* 371, 2477–2487 (2014).
  128. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* 371, 2488–2498 (2014).
  129. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* 559, 400–404 (2018).
  130. Steensma, D. P. *et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* 126, 9–16 (2015).
  131. Kumar, V. Diseases of the White Blood Cells, Lymph Nodes, Spleen and Thymus. in *Robbins and Cotran Pathologic Basis of Disease* (ed. Kumar, V.) 1450 (Elsevier Saunders, 2010).
  132. Valent, P. *et al.* Standards and impact of hematopathology in myelodysplastic syndromes (MDS). *Oncotarget* 1, 483–496 (2010).
  133. Wang, S. A. *et al.* Acute erythroid leukemia with <20% bone marrow blasts is clinically and biologically similar to myelodysplastic syndrome with excess blasts. *Mod. Pathol.* 29, 1221–1231 (2016).



152. Landgren, O. *et al.* Increased risks of polycythemia vera, essential thrombocythemia, and myelofibrosis among 24,577 first-degree relatives of 11,039 patients with myeloproliferative neoplasms in Sweden. *Blood* 112, 2199–2204 (2008).
153. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* 347, 78–81 (2015).
154. IARC. Most types of cancer not due to “bad luck” -IARC responds to scientific article claiming that environmental and lifestyle factors account for less than one third of cancers. *IARC Press Releases* 1–2 (2015).
155. Tomasetti, C. & Vogelstein, B. Response. *Science* 347, 729–731 (2015).
156. Ashford, N. A. *et al.* Cancer risk: role of environment. *Science* 347, 727 (2015).
157. Potter, J. D. & Prentice, R. L. Cancer risk: tumor excluded. *Science* 347, 727 (2015).
158. Golemis, E. A. *et al.* Molecular mechanisms of the preventable causes of cancer in the United States. *Genes Dev.* 32, 868–902 (2018).
159. Shiono, P. H., Chung, C. S. & Myriantopoulos, N. C. Preconception radiation, intrauterine diagnostic radiation, and childhood neoplasia. *J. Natl. Cancer Inst.* 65, 681–686 (1980).
160. Dickinson, H. O. & Parker, L. Leukaemia and non-Hodgkin’s lymphoma in children of male sellafield radiation workers. *Int. J. Cancer* 99, 437–444 (2002).
161. Nomura, T. Transgenerational effects of radiation and chemicals in mice and humans. *J. Radiat. Res.* 47 Suppl B, B83-97 (2006).
162. Barnett, L. B., Tyl, R. W., Shane, B. S., Shelby, M. D. & Lewis, S. E. Transmission of mutations in the lacI transgene to the offspring of ENU-treated Big Blue® male mice. *Environ. Mol. Mutagen.* 40, 251–257 (2002).
163. Kipen, H. M. & Wartenberg, D. Lymphohematopoietic Malignancies (Section 30.2). in *Textbook of Clinical Occupational and Environmental Medicine* (eds. Russi, M. B. et al.) 742–756 (Elsevier, 2005). doi:10.1016/B978-0-7216-8974-6.50034-8
164. Jung, M. *et al.* GATA2 deficiency and human hematopoietic development modeled using induced pluripotent stem cells. *Blood Adv.* 2, 3553–3565 (2018).
165. Mendizabal, A. M., Younes, N. & Levine, P. H. Geographic and income variations in age at diagnosis and incidence of chronic myeloid leukemia. *Int. J. Hematol.* 103, 70–78 (2016).
166. Richardson, D. B. Temporal variation in the association between benzene and leukemia mortality. *Environ. Health Perspect.* 116, 370–374 (2008).
167. Zeidan, A. M., Shallis, R. M., Wang, R., Davidoff, A. & Ma, X. Epidemiology of myelodysplastic syndromes: why characterizing the beast is a prerequisite to taming it. *Blood Rev.* 34, 1–15 (2019).
168. Hattis, D., Goble, R. & Chu, M. Age-related differences in susceptibility to carcinogenesis. II. Approaches for application and uncertainty analyses for individual genetically acting carcinogens. *Environ. Health Perspect.* 113, 509–516 (2005).
169. Mathews, J. D. *et al.* Cancer risk in 680000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. *BMJ* (2013). doi:10.1136/bmj.f2360
170. Veys, C. A. ABC of Work Related Disorders. *Br. Med. J.* 313, 615–619 (1996).
171. Filippini, T. *et al.* Association between outdoor air pollution and childhood leukemia: a systematic review and dose-response meta-analysis. *Environ. Health Perspect.* 127, 46002 (2019).
172. Shiels, M. S., Pfeiffer, R. M. & Engels, E. A. Age at cancer diagnosis among people with AIDS in the United States. *Ann. Intern. Med.* 153, 452–460 (2010).



173. Aziz, H. *et al.* Increased incidence of early onset colorectal cancer in Arizona: a comprehensive 15-year analysis of the Arizona Cancer Registry. *J. Gastrointest. Dig. Syst.* 5, (2015).
174. Bitchatchi, E., Kayser, K., Perelman, M. & Richter, E. D. E. D. Mesothelioma and asbestosis in a young woman following occupational asbestos exposure: short latency and long survival: Case Report. *Diagn. Pathol.* 5, 81 (2010).
175. Bitchatchi, E. Y., Levy, O., Richter, D., Fireman, E. & Baruch, R. Multi-system disease and various cancers in nonsmoking teachers: is chalk dust an occupational risk? in *Annals of the New York Academy of Sciences* 1076, 925–941 (Blackwell Publishing Inc, 2006).
176. Richter, E. D. *et al.* Cancer risks in naval divers with multiple exposures to carcinogens. *Environ. Health Perspect.* 111, 609–617 (2003).
177. White, M. C. *et al.* Age and cancer risk: a potentially modifiable relationship. *Am. J. Prev. Med.* 46, 1–16 (2014).
178. Leslie, K. D., Lynch, C. F. & Smith, E. M. Cancer. in *Maxcy-Rosenau-Last Public Health and Preventive Medicine* (eds. Wallace, R. B., Kohatsu, N. & Last, J. M.) 1047–1070 (McGraw-Hill Education/Medical, 2008). doi:10.1036/0071441980
179. Little, M. P. *et al.* Leukaemia and myeloid malignancy among people exposed to low doses (<100 mSv) of ionising radiation during childhood: a pooled analysis of nine historical cohort studies. *Lancet. Haematol.* 5, e346–e358 (2018).
180. Hourigan, C. S. & Savani, B. N. Acute leukemia risk associated with low dose radiation. *Lancet Haematol.* 5, e324–e325 (2018).
181. Shimon, I., Kneller, A. & Olchovsky, D. Chronic myeloid leukaemia following iodine-131 treatment for thyroid carcinoma: a report of two cases and review of the literature. *Clin. Endocrinol. (Oxf).* 43, 651–654 (1995).
182. Oluwasanjo, A., Pathak, R., Ukaigwe, A. & Alese, O. Therapy-related acute myeloid leukemia following radioactive iodine treatment for thyroid cancer. *Cancer Causes Control* 27, 143–146 (2015).
183. Guru Murthy, G. S. & Abedin, S. Myeloid malignancies after treatment for solid tumours. *Best Pract. Res. Clin. Haematol.* 32, 40–46 (2019).
184. Jephcote, C., Brown, D., Verbeek, T. & Mah, A. A systematic review and meta-analysis of haematological malignancies in residents living near petrochemical facilities. *Environ. Heal. A Glob. Access Sci. Source* 19, 1–18 (2020).
185. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *Benzene. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* 120, (International Agency for Research on Cancer, 2018).
186. Allegra, A. *et al.* Formaldehyde exposure and acute myeloid leukemia: a review of the literature. *Med.* 55, 1–9 (2019).
187. Mundt, K. A., Gentry, P. R., Dell, L. D., Rodricks, J. V. & Boffetta, P. Six years after the NRC review of EPA’s Draft IRIS Toxicological Review of Formaldehyde: regulatory implications of new science in evaluating formaldehyde leukemogenicity. *Regul. Toxicol. Pharmacol.* 92, 472–490 (2018).
188. Qin, L., Deng, H.-Y., Chen, S.-J. & Wei, W. Relationship between cigarette smoking and risk of chronic myeloid leukaemia: a meta-analysis of epidemiological studies. *Hematology* 22, 193–200 (2017).
189. Foucault, A. *et al.* Occupational pesticide exposure increases risk of acute myeloid leukemia: a meta-analysis of case–control studies including 3,955 cases and 9,948 controls. *Sci. Rep.* 11, 1–13 (2021).
190. Descatha, A., Jenabian, A., Conso, F. & Ameille, J. Occupational exposures and

- haematological malignancies: overview on human recent data. *Cancer Causes Control* 16, 939–953 (2005).
191. U.S.National Library Of Medicine. Haz-Map®. Leukemia. *Specialized Information Services* [<https://hazmap.nlm.nih.gov/category-details?id=167&table=tbl diseases>] (2018).
  192. Lutzmann, M. *et al.* MCM8- and MCM9 deficiencies cause lifelong increased hematopoietic DNA damage driving p53-dependent myeloid tumors. *Cell Rep.* 28, 2851–2865 (2019).
  193. Coltro, G. & Patnaik, M. M. Chronic myelomonocytic leukemia: insights into biology, prognostic factors, and treatment. *Curr. Oncol. Rep.* 21, (2019).
  194. Zimta, A. A., Tomuleasa, C., Sahnoune, I., Calin, G. A. & Berindan-Neagoe, I. Long non-coding RNAs in myeloid malignancies. *Front. Oncol.* 9, 1–19 (2019).
  195. Charrot, S., Armes, H., Rio-Machin, A. & Fitzgibbon, J. AML through the prism of molecular genetics. *Br. J. Haematol.* 188, 49–62 (2020).
  196. Zeisig, B. B., Kulasekararaj, A. G., Mufti, G. J. & Eric So, C. W. SnapShot: Acute myeloid leukemia. *Cancer Cell* 22, 698-698.e1 (2012).
  197. Greenberg, P. L. *et al.* Myelodysplastic syndromes, version 2.2017 Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* 15, 60–87 (2017).
  198. Baastrup Nordsborg, R., Meliker, J. R., Kjær Ersbøll, A., Jacquez, G. M. & Raaschou-Nielsen, O. Space-time clustering of non-Hodgkin lymphoma using residential histories in a Danish case-control study. *PLoS One* 8, (2013).
  199. Kingsley, B. S., Schmeichel, K. L. & Rubin, C. H. An update on cancer cluster activities at the Centers for Disease Control and Prevention. *Environ. Health Perspect.* 115, 165–171 (2007).
  200. Wasserstein, R. L. & Lazar, N. A. The ASA statement on p-values: context, process, and purpose. *Am. Stat.* 70, 129–133 (2016).
  201. Amrhein, V., Greenland, S. & Mcshane, B. Retire statistical significance. *Nature* 567, 305–307 (2019).
  202. Greenland, S. *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31, 337–350 (2016).
  203. Izquierdo Font, À., Marcos-Gragera, R., Vilardell Gill, M. L., Buxó Pujolràs, M. & Fuentes Fdz., J. *GCR 2012: El Càncer a Girona 2005-2006*. (Registre de Càncer de Girona, 2012).
  204. Catalanian Statistics Institute IDESCAT. (2020). Available at: <https://www.idescat.cat/>.
  205. Chen, R., Connelly, R. R. & Mantel, N. The efficiency of the sets and the cuscore techniques under biased baseline rates. *Stat. Med.* 16, 1401–11 (1997).
  206. Gail, M. H., Benichou, J. & Armitage, P. Cancer Registries. *Encyclopedia of Epidemiologic Methods* 124–138 (2000).
  207. Eayres, D. Analytical Tools For Public Health: Commonly Used Public Health Statistics and their Confidence Intervals. *Technical Briefing 3 of the Association of Public Health Observatories* [<https://wearchive.nationalarchives.gov.uk/20080906025804/http://www.apho.org.uk/resource/item.aspx?RID=48457>] 11 (2014).
  208. Reed, J. F. Better binomial confidence intervals. *J. Mod. Appl. Stat. Methods* 6, 153–161 (2007).
  209. R Core Team. R: A language and environment for statistical computing. (2019).
  210. *Cancer Incidence in Five Continents Volume IX. IARC Scientific Publication 160* 9, (IARC, 2007).
  211. *Cancer Incidence in Five Continents Vol. X. IARC Scientific Publication 164* 10, (IARC;IACR, 2014).



212. Cancer Incidence in Five Continents (Vol. IX): Indices of Data Quality /Leukaemia. Available at: [https://ci5.iarc.fr/CI5I-X/Pages/Table9q\\_sel.aspx](https://ci5.iarc.fr/CI5I-X/Pages/Table9q_sel.aspx). (Accessed: 1st July 2020)
213. Cancer Incidence in Five Continents (Vol. X): Indices of Data Quality /Leukaemia. Available at: [https://ci5.iarc.fr/CI5I-X/Pages/Table10q\\_sel.aspx](https://ci5.iarc.fr/CI5I-X/Pages/Table10q_sel.aspx). (Accessed: 1st July 2020)
214. National Institute of Geography. {Satellite view featuring aerial distance}. *Google Maps & Google Earth* (2019). Available at: <https://www.google.com/maps/@41.8363012,2.5213368,9321m/data=!3m1!1e3>. (Accessed: 9th December 2019)
215. Logeman, C. J. *et al.* Editorial: Cancer in small states – no small matter. *Cancer Epidemiol.* 50, Part B, 173–175 (2017).
216. Landrigan, P. J. *et al.* Pollution and children’s health. *Sci. Total Environ.* 650, 2389–2394 (2019).
217. Waller, L. A. Commentary: Regarding assessments of chance in investigations of ‘cluster series’. *International Journal of Epidemiology* 449–452 (2013). doi:10.1093/ije/dys238
218. Du, W. *et al.* Adverse drug reactions due to opioid analgesic use in New South Wales, Australia: a spatial-temporal analysis. *BMC Pharmacol. Toxicol.* 20, 55 (2019).
219. De Sario, M., Vecchi, S., Schifano, P. & Michelozzi, P. Revisione degli studi di cluster di leucemia infantile. *Epidemiol. Prev.* 40, 38–41 (2016).
220. Glass, D. C., Schnatter, A. R., Tang, G., Irons, R. D. & Rushton, L. Risk of myeloproliferative disease and chronic myeloid leukaemia following exposure to low-level benzene in a nested case–control study of petroleum workers. *Occup. Environ. Med.* 71, 266–274 (2014).
221. Working Group of the Academy of Medical Science. *Identifying the Environmental Causes of Disease: How Should we Decide What to Believe and When to Take Action?* (The Academy of Medical Science, 2007).
222. Amundadottir, L. T. *et al.* Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Med.* 1, e65 (2004).
223. Carroll, R. *et al.* Space-time variation of respiratory cancers in South Carolina: a flexible multivariate mixture modeling approach to risk estimation. *Ann. Epidemiol.* 27, 42–51 (2017).
224. Rohrbacher, M. & Hasford, J. Epidemiology and Etiology of Chronic Myeloid Leukemia. in *Neoplastic Diseases of the Blood* (eds. Wiernik, P. H., Dutcher, J. P. & Gertz, M. A.) 9–17 (Springer International Publishing, 2018). doi:10.1007/978-3-319-64263-5\_2
225. Copley, G. B. *et al.* Hospital-based case-control study of MDS subtypes and benzene exposure in Shanghai. *J. Occup. Environ. Med.* 59, 349–355 (2017).
226. Maynadié, M. *et al.* Twenty-five years of epidemiological recording on myeloid malignancies: data from the specialized registry of hematologic malignancies of côte d’or (Burgundy, France). *Haematologica* 96, 55–61 (2011).
227. Clegg, L. X., Feuer, E. J., Midthune, D. N., Fay, M. P. & Hankey, B. F. Impact of reporting delay and reporting error on cancer incidence rates and trends. *J. Natl. Cancer Inst.* 94, 1537–1545 (2002).
228. Catalanian Statistics Institute IDESCAT. *Estadística Demogràfica: La Població Estrangera 2004*. (Generalitat de Catalunya, 2007).
229. Alexander, F. E. & Cuzick, J. Methods for the assessment of disease clusters. in *Geographical and Environmental Epidemiology. Methods for small area studies*



- (eds. Elliott, P., Cuzick, J., English, D. & Stern, R.) 238–250 (Oxford University Press, WHO-Regional Office for Europe, 1996).
230. Temporal Clusters vs Voltaren vs Codein 43 years by Boolean searches by PubMed. *PubMed* (2020). Available at: [https://pubmed.ncbi.nlm.nih.gov/?term=%28%28diclofenac+%29OR+%28voltaren%29%29+NOT+%28adverse+effects%29+1977%2F01%2F01%3A2020%2F06%2F21%5Bdp%5D&sort=pubdate&sort\\_order=asc&size=200](https://pubmed.ncbi.nlm.nih.gov/?term=%28%28diclofenac+%29OR+%28voltaren%29%29+NOT+%28adverse+effects%29+1977%2F01%2F01%3A2020%2F06%2F21%5Bdp%5D&sort=pubdate&sort_order=asc&size=200). (Accessed: 21st June 2020)
  231. Mansfield, A., Stehr-Green, J. K. & Stehr-Green, P. A. Cluster Investigations of Non-Infectious Health Events. *North Carolina Cent. Public Heal. Prep.* 5, 1–8
  232. Ozonoff, D., Aschengrau, A. & Coogan, P. Cancer in the vicinity of a Department of Defense superfund site in Massachusetts. *Toxicol. Ind. Health* 10, 119–141 (1994).
  233. Yan, P. & Clayton, M. K. A cluster model for space-time disease counts. *Stat. Med.* 25, 867–881 (2006).
  234. Jacquez, G. M. *et al.* Global, local and focused geographic clustering for case-control data with residential histories. *Environ. Heal. A Glob. Access Sci. Source* 4, 1–19 (2005).
  235. Sego, L. H., Reynolds, M. R. & Woodall, W. H. Risk-adjusted monitoring of survival times. *Stat. Med.* 28, 1386–1401 (2009).
  236. Grigg, O. A. & Farewell, V. T. A risk-adjusted sets method for monitoring adverse medical outcomes. *Stat. Med.* 23, 1593–1602 (2004).
  237. Radaelli, G. Using the Cuscore technique in the surveillance of rare health events. *J. Appl. Stat.* 19, 75–81 (1992).
  238. Smith, R. One Bristol, but there could have been many. *Bmj* 323, 179–180 (2001).
  239. Aylin, P. *et al.* Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984-96: was Bristol an outlier? *Lancet* 358, 181–187 (2001).
  240. Munro, A. J. Comparative cancer survival in European countries. *Br. Med. Bull.* 110, 5–22 (2014).
  241. Crosignani, P. Re: False-positive results in cancer epidemiology: a plea for epistemological modesty. *J. Natl. Cancer Inst.* 101, 212–3; author reply 213-4 (2009).
  242. Clapp, R. W. & Kriebel, D. Re: False-positive results in cancer epidemiology: a plea for epistemological modesty. *J. Natl. Cancer Inst.* 101, 211–2; author reply 213-4 (2009).
  243. Hauptmann, M. & Ronckers, C. M. RE: A further plea for adherence to the principles underlying science in general and the epidemiologic enterprise in particular. *Int. J. Epidemiol.* 39, 1677–9; author reply 1679-80 (2010).
  244. Hill, A. B. President’s address: The Environment and Disease: Association or Causation. *Proc. R. Soc. Med.* 7–12 (1965).
  245. Nicholson, W. J. & Landrigan, P. J. Quantitative assessment of lives lost due to delay in the regulation of occupational exposure to benzene. *Environ. Health Perspect.* 82, 185–188 (1989).
  246. Driesen, D. M. Is cost-benefit analysis neutral? *Univ. Color. Law Rev.* 77, 1–74 (2006).
  247. Michaels, D. & Monforton, C. Manufacturing uncertainty: contested science and the protection of the public’s health and environment. *Am. J. Public Health* 95, 39–48 (2005).



## 9. APPENDIX

### APPENDIX 1

Article: Detection and estimation of the increasing trend of cancer incidence in relatively small populations



Cancer Epidemiology 50 (2017) 207–213

Contents lists available at ScienceDirect

**Cancer Epidemiology**

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: [www.cancerepidemiology.net](http://www.cancerepidemiology.net)





---

## Detection and estimation of the increasing trend of cancer incidence in relatively small populations

Rina Chen<sup>a,\*</sup>, Enrique Y. Bitchatchi<sup>b</sup>

<sup>a</sup>BioForum, Applied Knowledge Center, Ness-Ziona, Israel  
<sup>b</sup>University of Gerona, Gerona, Spain



---

**ARTICLE INFO**

**Article history:**  
 Received 27 October 2016  
 Received in revised form 2 April 2017  
 Accepted 6 April 2017

**Keywords:**  
 Joinpoint regression  
 Trend in cancer incidence  
 Temporal clustering  
 Cuscore test  
 Relative interval (RI)  
 q-interval

**ABSTRACT**

**Background:** Detection and estimation of trends in cancer incidence rates are commonly achieved by fitting standardized rates to a joinpoint log-linear regression. The efficiency of this approach is inadequate when applied to a relatively low levels of incidence. We compared that approach with the Cuscore test with respect to detecting a log-linear increasing trend of chronic myelomonocytic leukemia (CMML) in datasets simulated to match a province of about 700,000 inhabitants.

**Methods:** For better efficiency, we replaced the standardized rate as the dependent variable with a continuous statistic that reflects the inverse of the standardized incidence ratio (SIR). Both procedures were applied to datasets simulated to match published results in the Girona Province of Spain. We also present the use of the q-interval in displaying the temporal pattern of the events. This approach is demonstrated by analyses of CMML diagnoses in Girona County (1994–2008).

**Results:** The Cuscore was clearly more efficient than regression in detecting the simulated trend. The relative efficiency of the Cuscore is likely to be maintained in even higher levels of incidence. The use of graphical displays in providing clues regarding interpretation of the results is demonstrated.

**Conclusions:** The Cuscore test coupled with visual inspection of the temporal pattern of the events seems to be more efficient than regression analysis in detecting and interpreting data suspected to be at elevated risk. A confirmatory analysis is expected to weed out 75% of the superfluous significant results.

© 2017 Elsevier Ltd. All rights reserved.

---

**1. Introduction**

Prediction of cancer incidence in a given population is often based on fitting a joinpoint regression model [1–4] to the logged standardized rates. This model assumes several consecutive linear trends (connected at joined points). The predicted incidence is based on the estimated annual percentage change (APC) derived from the slope of the last part of the joinpoint regression. Although it is well known that the efficiency of this procedure is inadequate when applied to small numbers of incident cases [3,5], it has been applied in such situations even when no incident cases were observed [4,5].

That inefficiency may be related to the difficulty in complying with linearity and with other restricted conditions underlying regression. In contrast, the efficiency of temporal clustering techniques is based only on appearance of clustering

among some consecutive cases. Such clustering is expected under any form of increased rate (linear or not). Based on that reasoning, we compared the efficiency of the Cuscore test [6–9] with that of regression in detecting a log-linear increasing rate in data of small incident numbers. The datasets were simulated to match published results of chronic myelomonocytic leukemia (CMML; ICD-O-3 code: 9945/7) in Girona province, Spain [4].

We replaced the usual dependent variable in the regression analyses with the RI (relative interval) statistic. RI measures the time intervening between two consecutive events where each event includes a predefined number (r) of consecutive cases. Being a continuous positive variable, it is somewhat more stable than any adjusted rate (as the random variable of any adjusted rate is a function of the annual observed number of cases) and bypasses the need to deal with zero observed cases in some years.

---

\* Correspondence to: 38/33 Hanassi St., Tel. Aviv 69206, Israel.  
 E-mail addresses: [rinachen@netvision.net.il](mailto:rinachen@netvision.net.il) (R. Chen),  
[eboccupenviron@gmail.com](mailto:eboccupenviron@gmail.com) (E.Y. Bitchatchi).

<http://dx.doi.org/10.1016/j.canep.2017.04.005>  
 1877-7821/© 2017 Elsevier Ltd. All rights reserved.

## 2. Data

### 2.1. Simulated datasets

Simulation of 300 datasets was carried out assuming 61 cases diagnosed in a 15-year stable population with respect to size and risk factors. Under the geometric series with an annual increase of 3.3%, the number of cases in the first year was calculated to be  $n_1 = 3.21$ .

The simulation was carried out for 60 cases, assuming that the expected number of new cases in year  $t$  is:  $E(X_t) = 3.21 \cdot 1.033^{t-1}$ . Accordingly, the expected interval between consecutive cases is  $E(W_t) = 12/E(X_t)$  months. Assuming exponential distribution, the number of months between consecutive diagnoses was randomly allocated (using STATA [10]). When an interval extended over 2 years, we used a rule (described below) whereby the expected interval was determined as either  $E(W_t)$  or  $E(W_{t+1})$ . It should be noted that either choice leads to a biased allocated time interval. However, for data covering 15 years, the bias will be in intervals of 14 (out of 60) cases at most, and will affect the waiting time of only  $1/r$  cases of the relevant event. In order to minimize bias, the choice between  $E(W_t)$  and  $E(W_{t+1})$  was made such that most of the interval is allocated under the correct  $E(W_t)$ . By our rule of thumb, the allocated time interval was according to  $E(W_{t+1})$  if the time remaining in year  $t$  after the diagnosis date of the previous case was less than  $E(W_t)/4$ .

The analyses are based on the recorded time of events where each event includes  $r$  consecutive cases. The three  $r$  values are: 3, 4 and 5. We grouped the data into three  $r$  groups (each with 100 datasets) according to the size of  $r$ .

### 2.2. CMML cases in Girona County (1994–2008)

Our approach is demonstrated using real data recorded during a 15-year period in Girona County.

#### 2.2.1. Girona County Region (the Central Comarca of the Province of Girona)

Girona County constitutes about one-fourth of Girona province's population and half of its latitude. The community of the county is better off than the community of Catalonia *en bloc* with respect to economic welfare and healthcare availability. In general, residential communities are quite stable (in size and profile) over the 15-year period. However, an influx of young people of working age began in the mid-1990s. The possible effect of that immigration on the age profile of our analyzed data was found to be negligibly small [11,12].

#### 2.2.2. Reference population

The best affordable a priori reference population was extracted as an aggregate of the annual counts over 36 age and gender strata and 221 municipalities of Girona Province. Catalanian Statistics Institute (IDESCAT) strata counts exist only for the last 10 years (1999–2008). However, municipal census could be accessed directly. The ultimate dataset originated from municipalities' census apiece plus smoothing splines generalized linear model (GLM) including Poisson response. This yielded a 15-year aggregated reference for 1994–2008.

#### 2.2.3. Annual expected number of cases

Specific age and gender incidence rates were assumed to be the rates observed in the Girona Province during the 15-year period from 1994 to 2008. Denoting by  $R_{g,t}$  the age- and gender-specific rate in the reference population, and by  $N_{s,t}$  the relevant group size in Girona County in year  $t$ , the expected number of new cases in year  $t$  is:  $E(X_t) = \sum R_{g,t} \cdot N_{s,t}$ .

## 3. Methods

### 3.1. Test statistics

#### 3.1.1. The $RI$

The  $RI$  statistic is defined as  $w/E(W)$ , where  $w$  is the observed number of months intervening between two consecutive events and  $E(W)$  is the expected number of months between two consecutive diagnoses. Accordingly, the length of  $RI$  is simply the expected number of cases during  $w$  months. It can easily be updated with respect to temporal changes in the population's profile by updating annually the expected number of cases. Thus,  $RI$  is the waiting time until the event, measured by the expected number of cases regardless of the current  $E(W)$  length. As such its distribution is gamma and its mean is the event size  $r$ . The fact that the mean time until the event equals the event size is clear. This is so since the expected time until a single case is  $E(W)$  months, hence the expected time until  $r$  cases is  $r \cdot E(W)$ , thus  $RI = r \cdot E(W)/E(W) = r$ . Practical details of  $RI$  calculation are demonstrated in 4.2.

It is interesting to note that when the annual expected number of cases is  $r$ ,  $RI/r$  is the inverse of SIR, since  $RI/r = \exp(\text{obs}) \approx 1/\text{SIR}$ . The difference between the two measures is the random variable. It is the observed number in SIR and the expected number of cases in  $RI$ .

Based on this, and in order to comply with the common practice in which trend analysis is based on the annual incidence, we suggest that  $r$  is defined as the upper integer of the annual expected number of cases at baseline. It is quite likely – even under an increasing trend – that the number of cases expected annually is still close to  $r$ . In our simulated data, analyses are based on  $r = 4$ . We also analyzed data in which the event included three or five cases. Results of these two  $r$  values provided better insight regarding the relative efficiency of the two tests.

The easy accommodation of  $RI$  under changing conditions, and the fact that its gamma distribution depends only on  $r$ , enabled the derivation of several procedures aimed at detection and interpretation of the increased rate of cancer diagnoses [6–9,13–18].

#### 3.1.2. The $q$ -interval

The  $q$ -interval [17] is defined as the a-priori probability that the waiting time until the event is longer than that observed. Namely, it is the a-priori probability that the  $r^{\text{th}}$  case of an event is diagnosed after the observed  $RI$ . As a gamma distributed variable, the  $q$ -interval can also be calculated under the Poisson distribution. Under the Poisson it is defined as the probability that no more than  $r-1$  cases are observed during an interval in which  $RI$  cases should be expected. For example, suppose that  $r = 4$ ,  $E(W) = 3$  months and  $w = 16$  months. Namely, four cases were observed during a period in which ( $RI = 16/3 = 5.33$ ) 5.33 cases should be expected. According to the Poisson distribution, the probability that no more than three cases are observed during  $RI = 5.33$  is the  $q$ -interval = 0.222.

Although the  $q$ -interval is calculated as a probability value, it is actually a random variable derived from an observed random event ( $RI$ ). It is a cumulative probability of a continuous random variable; as such, its distribution is uniform over 0–1 [19]. Accordingly, under stable conditions, its expected value is 0.5 (for any  $r$  value) and  $>0.5$  under elevated incidence. Based on that distribution we can use the  $q$ -interval in graphical display of the temporal pattern of the events.

### 3.2. Analyses

Both regression and Cuscore procedures were applied to each of the simulated datasets. The relative efficiency of the two procedures was tested for significance (two-tailed) by applying McNemar's test [20].

**Table 1**  
 Critical values ( $P_{crit}$  and  $RI_{crit}$ ) by number of events ( $S$ ) and number of cases in each event ( $r$ ).

S	$P_{crit}$	$RI_{crit}$						
		r=1	r=2	r=3	r=4	r=5	r=6	r=7
5	0.549	0.797	1.840	2.878	3.911	4.940	5.967	6.991
6	0.507	0.707	1.701	2.703	3.706	4.709	5.712	6.715
7	0.446	0.590	1.511	2.460	3.420	4.385	5.355	6.327
8	0.421	0.546	1.437	2.365	3.306	4.256	5.211	6.170
9	0.394	0.501	1.359	2.263	3.184	4.117	5.057	6.002
10	0.377	0.473	1.311	2.199	3.108	4.030	4.960	5.896
11	0.362	0.449	1.269	2.143	3.041	3.953	4.874	5.802
12	0.351	0.432	1.238	2.102	2.992	3.897	4.811	5.734
13	0.341	0.417	1.210	2.065	2.947	3.845	4.754	5.671
14	0.333	0.405	1.188	2.036	2.912	3.804	4.708	5.621
15	0.326	0.395	1.169	2.010	2.880	3.768	4.668	5.576
16	0.319	0.384	1.149	1.984	2.849	3.732	4.628	5.532
17	0.313	0.375	1.133	1.962	2.822	3.701	4.593	5.494
18	0.308	0.368	1.119	1.943	2.800	3.675	4.564	5.462
19	0.304	0.362	1.108	1.929	2.782	3.654	4.540	5.436
20	0.299	0.355	1.095	1.910	2.759	3.628	4.511	5.404

$P_{crit}$  and  $RI_{crit}$  are critical values of  $P$  and of  $RI$ .  $P$  is the null probability for a shorter time interval than that observed between consecutive events.  $RI$  is the expected number of cases during an observed time interval. The value of  $RI_{crit}$  (or  $P_{crit}$ ) is needed for the Cuscore test at 5% significance.

Analysis of real data of Girona County was based on the Cuscore test, supplemented by graphical presentation of the cumulative  $q$ -interval.

### 3.2.1. Regression

We replaced the usual dependent variable in the regression analyses with the  $RI$ . Being a continuous positive variable, it is somewhat more stable than any adjusted rate (as the random variable of any adjusted rate is a function of the observed number of cases).

The joinpoint regression reduces to the standard linear regression when the trend is continuous over the entire data. Accordingly, the regression model was applied to  $\ln(RI)$  over consecutive events. The significance of the slope ( $b$ ) was determined at the 5% level by a one-tailed test. The mean and 95% confidence limits (95%CL) of APC were evaluated for each  $r$ -group, as:  $APC = 100 \cdot (1 - \exp(\bar{b}))$  where  $\bar{b}$  is the mean of  $b$  over all datasets of the group. The 95%CLs were each estimated by replacing  $\bar{b}$  with the corresponding CL.

### 3.2.2. The Cuscore test

The Cuscore test [6–9] is aimed at detecting temporal clustering of events. For that test, each  $RI$  is defined as either “short” or “long” according to a given critical value ( $RI_{crit}$ ). Technically, the test is based on an accumulated score (Cuscore) of “short”  $RI$ s. At the start, Cuscore=0, and with each event it either increases by 1 (if  $RI \leq RI_{crit}$ ) or decreases by 1 (if  $RI > RI_{crit}$ ). Departing from that rule, the Cuscore remains 0 when the next interval is “long”. A significantly increased rate is declared if the Cuscore=5 at any point during the sequential analyses.

The  $RI_{crit}$  values were derived from  $P_{crit}$ , where  $P_{crit} = Pr(RI \leq RI_{crit})$ . The  $P_{crit}$  values were evaluated [9] by Markov-Chain so that the probability for Cuscore=5 among  $S$  events is 0.05. (Details regarding the Markov-Chain approach are presented in Appendix A).  $RI_{crit}$  corresponding to the relevant  $P_{crit}$  was evaluated under gamma distribution [7,8] (see Appendix A). Table 1 presents  $P_{crit}$  and  $RI_{crit}$  values by  $S$  and  $r$  for 5% (one-sided) significance.

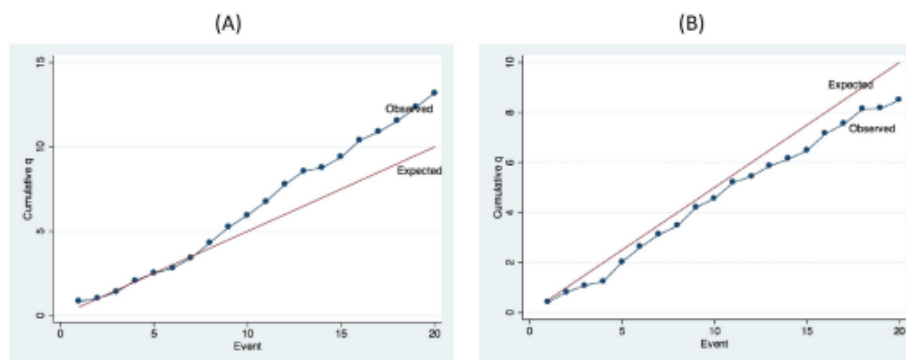
### 3.2.3. Graphical presentation of trends

A  $q$ -interval is expected to be 0.5 under stable conditions and  $>0.5$  under elevated rates. In view of the small counts of cases and assuming a moderately increasing trend induced by a new exposure, the increasing  $q$ -intervals may not be apparent as would be the accumulative  $q$ -intervals ( $Sq$ ) [17].

Clearly, the expected slope of the  $Sq$  curve is 0.5 under stable conditions and  $>0.5$  under elevated rates. An “eyeball” inspection of the depicted observed and expected  $Sq$  may be useful in detecting an increased rate. However, it should be noted that the slope of the line may be misleading because each  $Sq$  depends on the preceding  $q$ -intervals.

Fig. 1A demonstrates the value of the  $Sq$  curve in the detection and interpretation of data. It presents the  $Sq$  of one set of our simulated data ( $APC = 3.3\%$  &  $r = 3$ ). Fig. 1B demonstrates the results of data simulated under  $APC = 0$  for 60 cases (60) and  $r = 3$ .

The slopes in Fig. 1A indicate elevation from the 7th event (at the 8th year), where the slope seems to be larger than expected. The curve bears no indication that the increasing incidence started much earlier (at the 2nd year), apparently because of the very small incident counts in the early years. Also, it cannot possibly provide an indication of the fact that the logged rate followed a linear curve. The conclusion that the rate has been increasing since at least as early as year 8 may be of value in considering the



**Fig. 1.** Observed and expected cumulative  $q$ -intervals of two datasets simulated under an increasing trend (A) and under stable conditions (B). The  $q$ -interval is calculated for each event (that includes  $r = 3$  cases).

**Table 2**

Mean (and 95% CLs) of duration (years) and of annual percentage change (APC) in simulated datasets, by event size (*r*).

No. of cases in an event ( <i>r</i> )	Number of datasets	Duration (years)	APC
3	100	15.03 (14.76; 15.31)	2.2 (1.7; 2.6)
4	100	14.57 (14.26; 14.87)	3.0 (2.4; 3.6)
5	100	14.93 (14.63; 15.24)	3.8 (3.1; 4.6)
All	300	14.85 (14.68; 15.02)	

Duration is the time during which 60 cases are diagnosed. APC: annual percentage change.

responsible exposure [17]. It is interesting to note that, in that example, the graphical presentation was more sensitive than the formal analyses, as both the regression and the Cuscore analyses of that dataset were not significant.

In the general, eyeball inspection of *Sq* may reflect several possible interpretations of an elevated rate (see Discussion).

#### 4. Results

##### 4.1. Simulated datasets

Table 2 presents for each group of datasets the means (and 95% CLs) of APCs and the time span (i.e., the time period during which the 60 cases were observed). The span is expected to be the same in each of the *r*-groups, since the simulation of all the 300 sets were carried out under the same assumptions. According to the assumed geometric series, the expected duration for the 60 incident cases is 14.80 years (15.0 years for 61 cases). Indeed, the mean duration in each of the three *r*-groups, as well as that over the 300 sets (14.85 years), is quite close to that expected.

The mean APC increased with *r* from 2.2 to 3.8. For *r* = 3, APC is clearly an underestimate, as the upper CL (2.6%) is much below that expected (3.3%). Two factors may affect the slope of the regression line when *r* is small. One factor is the stability of *RI* (that increases with *r*). The other, which is the main factor, is related to the fact that the regression analyses were over consecutive events rather than over years. This may lead to biased APC estimates. In our simulated data, underestimated values are expected when *r* = 3 and overestimated values when *r* = 5. The reasoning behind these statements is presented in the Discussion section.

**Table 3**

Number of significant results by method of analysis in simulated datasets, using *r* = 3 in the analyses (where *r* = number of cases included in each event).

Cuscore	Regression		Total
	Significant	Non-significant	
Significant	4	12	16
Non-significant	16	68	84
Total	20	80	100

**Table 4**

Number of significant results by method of analysis in simulated datasets, using *r* = 4 in the analyses (where *r* = number of cases included in each event).

Cuscore	Regression		Total
	Significant	Non-significant	
Significant	13	26	39
Non-significant	9	52	61
Total	22	78	100

**Table 5**

Number of significant results by method of analysis in simulated datasets, using *r* = 5 in the analyses (where *r* = number of cases included in each event).

Cuscore	Regression		Total
	Significant	Non-significant	
Significant	13	29	42
Non-significant	12	46	58
Total	25	75	100

Tables 3–5 present the number of significant results by each method for each *r* group. Both procedures gain in efficiency with increased *r*, especially the Cuscore.

Relative to the regression procedure, the number of significant results is smaller (albeit not significantly so) by the Cuscore for *r* = 3 and clearly larger and highly significant ( $P < 0.01$ ) for *r* > 3. The largest difference between the two procedures is for *r* = 5, where significance was obtained for 42% by the Cuscore as compared to 25% by regression.

##### 4.2. CMML data in Girona County

Table 6 presents the expected and observed number of CMML cases in each year from 1994 to 2008. The SMR (standardized morbidity ratio) over the 15-year period was  $(19/14.56) = 1.30$  ( $P < 0.15$ ) and (assuming Poisson distribution) its 95% confidence limits are: 0.64; 1.66.

The Cuscore analysis was based on *r* = 1. Accordingly, there are 19 observed intervals plus the censored interval preceding the unobserved diagnosis date of the 20th case. For 19 sequential analyses,  $RI_{crit} = 0.362$ .

Table 7 presents for each case: the date of diagnosis, the *RI*, the Cuscore value and the *q*-interval. Fig. 2 presents the cumulative *q*-intervals. We use these data to demonstrate the calculation of the *RI*s using the time interval preceding case 1.

The *RI* of the first case is the number expected in 1994+ that expected in 2 months and 7 days of 1995. Assuming 30.5 days in a month, we get:

$$RI_1 = 1.023 + ((2 + (7/30.5)) * (1.011/12)) = 1.023 + 0.188 = 1.211.$$

Luckily, the data in Table 7 include the exact date. When we know only the year of the diagnosis we decide on the month according to the number of diagnoses in that year. For example, if only one case is diagnosed in year *t* than we assume that it happened in June 30. If there two diagnoses at year *t* we determine their dates in the first and second third of the 12 months. Real examples are presented in Ref. [17].

**Table 6**

Expected and observed number of chronic myelomonocytic leukemia (CMML) cases in Girona County (1994–2008) by year.

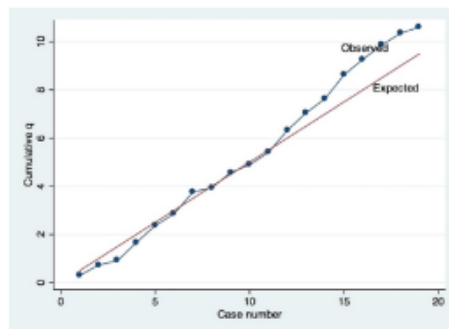
Year	Observed	Expected
1994	0	1.023
1995	1	1.011
1996	1	1.005
1997	2	1.006
1998	1	1.009
1999	2	0.905
2000	0	0.911
2001	2	0.881
2002	1	0.974
2003	3	0.970
2004	3	1.045
2005	1	1.059
2006	1	1.031
2007	1	0.851
2008	0	0.874
Total	19	14.555

**Table 7**  
 Diagnosis dates of chronic myelomonocytic leukemia (CMML) cases in Girona County (1994–2008) and the associated *R*, *q*-Interval, and Cuscore.

Case number	Age gender	Date	Interval		Score
			<i>R</i>	<i>q</i>	
1	72 m	07/03/1995	1.211	0.296	0
2	85 m	02/01/1996	0.829	0.437	0
3	92 m	19/09/1997	1.722	0.179	0
4	89 f	29/12/1997	0.278	0.757	1
5	80 f	06/05/1998	0.357	0.700	2
6	66 f	20/01/1999	0.706	0.494	1
7	90 m	28/02/1999	0.101	0.904	2
8	78 m	24/01/2001	1.724	0.178	1
9	75 m	06/08/2001	0.471	0.624	0
10	78 m	22/10/2002	1.142	0.319	0
11	82 m	26/06/2003	0.658	0.518	0
12	77 m	28/07/2003	0.086	0.918	1
13	75 m	26/11/2003	0.318	0.728	2
14	90 f	04/06/2004	0.539	0.583	1
15	76 f	07/06/2004	0.009	0.991	2
16	92 m	10/11/2004	0.444	0.641	1
17	83 m	05/05/2005	0.513	0.599	0
18	83 f	11/01/2006	0.723	0.485	0
19	74 m	31/05/2007	1.355	0.258	0
20	-	>31/12/2008	>1.370	<0.254	0

The time intervals between cases are measured in terms of *R* (expected number of cases) and *q*-interval (the null probability for a longer interval than that observed). The score is the Cuscore sequential value.

The results of the Cuscore presented in Table 7 are clearly not significant. However, the *q*-intervals vaguely indicate elevated rates after 2002, when *q*-interval >0.5 is observed for each of six cases in a sequence (see cases 12–17 in Table 7 and in Fig. 2). However, the last three intervals (including the censored 20th) contradict that indication. In fact, these intervals reflect a declining incidence rate. This seeming contradiction between the results by the Cuscore test and the impression obtained from the *q*-intervals may be explained by any one of several possibilities, including chance occurrence, inadequate power of the Cuscore in the detection of a rather small cluster embedded in the data, and introduction of a new diagnostic procedure that enables diagnosis at an earlier stage.



**Fig. 2.** Cumulative *q*-intervals of CMML cases in Girona County from 1994 to 2008. The Cumulative *q*-intervals value of case *i* is the sum of the *q*-intervals of all cases up to (and including) case *i*. The *q*-interval of each case is the null expected probability for a longer time than that elapsed since the diagnosis of the previous case.

## 5. Discussion

Our main interest is focused on the results of the  $r=4$  group. Results with the other two  $r$  values provide some insight regarding the relative efficiency of the two procedures under comparison.

For technical reasons (discussed below) the efficiency of both procedures are weakened when  $r \leq E(x_0)$ . Our interest should therefore be focused on results observed in data analyzed using  $r > 3$ . In the two groups of interest, the regression approach was significantly less efficient than the Cuscore. As explained below for the  $r=5$  group the slope of the regression line likely overestimated in data of  $r=5$ , hence lending support to the conclusion that the relative higher efficiency of the Cuscore is detecting a real increasing rate in similar data.

Because of several sources of variability, the efficiency of regression analysis in a similar or even larger real dataset is likely to be even poorer. These sources attenuate the slope of the regression line and might also break the hypothetical continuous linearity (which is only partly taken care of by the jointpoint regression). Since the Cuscore does not depend on linearity, its efficiency is expected to be higher than that of the regression in many situations.

There are some issues inherent in our specific simulation and analysis that may have attenuated the efficiency of the regression approach when  $r=3$ , leading to overestimation of the APC when  $r=5$ . These issues are related to the fact that the regression line was fitted over the sequential order of events rather than over years. In natural situations, when the increasing incidence is continuous, the size of  $r$  should not affect the slope of the regression over the events. This is not the situation in our simulation procedure, in which the expected number of cases was constant over each year  $t$  and jumped by 3.3% in  $t+1$ . Accordingly, two types of bias in *R* will occur. One occurs when two events are allocated to the same year, the other when the allocated interval of one case spreads over 2 consecutive years. The first is likely to occur for  $r=3$ , while the other may occur for  $r=5$ . Under both situations, the local slope of the regression line will be shallower during either the early part of the line (when  $r=5$ ) or during the later part ( $r=3$ ). Accordingly, the slope of the entire line will be underestimated when  $r=3$  and somewhat overestimated when  $r=5$ . A moderate effect is expected when  $r=5$ , since one or, at most, two intervals of five cases may be slightly affected.

Thus, our study indicates that the efficiency of the regression approach is quite poor even under reinforcing conditions (i.e., overestimated slope when  $r=5$ ) and that of the Cuscore test is better but still inadequate.

The low efficiency of both procedures is an inherent issue in detecting elevated incidences of cancer [6,15,17]. This low efficiency is coupled with an inflated significance level [6,17,21] that evolves from the huge and unknown sample space (chance occurrence may be observed in so many cancer diseases, periods, communities, etc.). So much so that Coory and Jordan [21] suggested avoiding the use of *P* values in investigations of clusters. While we agree with Coory and Jordan that the final decision should be based on results from epidemiologic investigation, we also agree with Assunção's response to that paper [22], in which he concludes that: "... statistical evidence is only one ingredient on which to base decision-making in cluster reports".

Our suggested approach goes somewhat beyond providing a *P* value. In addition to the test, it includes a display of the temporal pattern of the incidence and a confirmatory analysis [15] (that weeds out 75% of the superfluous significant results). These procedures allow some insights into the data and provide grounds for confirmation or denial of an exposure effect as well as clues needed for further investigations, including exposure assessment (as suggested by Coory and Jordan).

The pattern of the  $q$ -intervals displayed by the cumulative values may be useful in providing clues regarding the four possible interpretations of an observed temporal clustering. These are described below:

1. Incidental occurrence – the  $q$ -intervals in the vicinity of a detected clustering fluctuate around 0.5.
2. Inadequate reference population –  $q$ -intervals that tend to be larger (or smaller) than 0.5 over the entire study. A pattern showing larger values may also indicate that the studied community has been exposed to a carcinogen over a long time.
3. Enhanced diagnoses (caused by introduction of a procedure, instrument, etc.) – a sequence of  $q$ -intervals  $>0.5$ , followed by a sequence of  $q$ -intervals  $<0.5$ , then returning to fluctuate around 0.5. However, the available data may not yet show all the three phases.
4. A true clustering induced by new exposure to a carcinogen – in a residential community:  $q$ -intervals  $>0.5$  continually observed beyond the detected cluster. In a workplace or another community with low turnover, a pattern similar to that of enhanced diagnoses is expected. The decline of the incidence (i.e.,  $q$ -intervals  $<0.5$ ) is related to the fact that only a few individuals are likely to respond to any specific exposure, and for some of them the exposure enhances an existing carcinogenic process. Thus, eventually “at-risk” persons are all “resistant” to the relevant risk factors.

Following detection of an increasing trend by our suggested procedure, the APC can be estimated according to log-linear trend assumption. This assumption complies with that underlying the log-joinpoint regression application, and also with the common practice in which the distributions of both the latent period [23] and of the dose–response function are log-normal [24]. In a way, this contradicts Breslow and Day [25], who maintain that the actual distribution of the latency [p. 264] may not be log-normal and that the logarithmic dose–response function may not represent the function when the dose is accumulated over time [p. 181]. However, although the log-normal distribution may not be blindly used for in-depth studies, it may provide an approximate estimate needed for the APC.

A point of interest is the possible effect of inaccurate RI values on the results of the Cuscore test [26]. Such inaccuracies may be related to several factors (e.g., inaccuracy in the estimated expected number of cases for 1 or more years, or to unknown exact date of diagnoses, etc.). The effect on the significance of value may be important only when the relevant RIs are large and goes to the same direction, and also when the true values of consecutive RI values are close to the cut-point criterion for “short” RI. This is very likely to occur, since each RI is the expected number of cases summed over (age and gender) groups and years.

In conclusion: Our suggested analysis is based on four consecutive procedures. These include: the **Cuscore** test; **display** of the temporal pattern of the events; and **confirmatory** test applied (if possible) to data observed subsequently to a significant result. These three procedures enable weeding out of most of the superfluous significant results as well as enabling the detection of an increased rate even when results are merely on the verge of significance. The fourth procedure is **estimate of APC** by a simple calculation based on log-linear assumption.

#### Conflict of interest

The study was not financially supported. Neither one of the two authors was involved in any way that may create conflict of interests.

#### Contribution of authors

RC was responsible to the conception, design and statistical methodology (including simulation). EB was responsible for acquisition, quality control and analyses of the data in Girona. Both authors are responsible to manuscript preparation.

#### Appendix A.

The chance probability for a significant result (i.e. Cuscore = 5) within 5 observed events is evaluated using the Markov–Chain approach with  $n+1$  possible states. In general Markov–Chain represents a process where the move from one state to an adjacent other depends only on the current state. In the Cuscore test the states represent the current value of the Cuscore. Accordingly, there are six states (0, 1, 2, 3, 4, 5). After each event the Cuscore will move either to the next or to the previous event (except for state 0). Thus, for example, suppose that the Cuscore state is at state 2, following the next event the scheme will be either at state 3 (with probability  $P_C$ ) or at state 1 (with probability  $1-P_C$ ). Significance is declared if the scheme reaches state 5.

In order to calculate the probability for a significant result we use the transition matrix and the probability distribution vector. The transition matrix T is:

$$T = \begin{matrix} & q & p & 0 & 0 & 0 & 0 \\ & q & 0 & p & 0 & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 & 0 \\ 0 & 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Each entry in the matrix represents the probability that the scheme moves from state (row)  $i$  to state (column)  $j$ . The probability distribution vector represents the probability distribution of the scores after each event. Starting from the second event, the initial probability distribution vector is:  $(q \ p \ 0 \ 0 \ 0)$ . Namely after the first event the scheme is at state 0 with probability  $q$  or at state 1 with probability  $p$ . The product of this vector and the matrix T is the probability distribution after the second event. In general, the probability distribution after event  $k$  is obtained by the product of the distribution vector after event  $k-1$  and T. Results are significant when the scheme reaches state 5.

The probability that the Cuscore test arrives at state 5 during 5 events is 0.05 if  $p$  and  $q$  in the matrix and vector are replaced by  $P_C$  and  $1-P_C$  presented in Table 1. For ease of computations, Table 1 presents also the  $RI_{crit}$  values corresponding to each  $P_C$  and  $r$ . The correspondence of  $P_C$  and  $RI_{crit}$  is demonstrated by an example: suppose that  $S=15$  then  $P_C=0.326$ ; for  $r=4$   $RI_{crit}=2.880$ . This is based on the assumption that we predefined a period in which 2.88 cases are expected. Under the Poisson distribution the probability that the fourth case is not observed during that period is 0.326.

#### Appendix B.

When the data over the later period of the study indicate an increasing trend of the logged SIR but no adequate regression model can be fitted, we can estimate the APC under two assumptions: stability in the size and profile of the population (so that the null expected number of annual cases is approximately constant) and a constant APC. For that estimate, we first determine the period during which the trend seems to be increasing, then split the number of years by half. The total number of cases in each half is the sum of a geometric series with the same APC. Denoting by T1 and T2 the observed number of cases in each half, the following ratio between the two sums can be observed:





## APPENDIX 2

### **The trade-off false positive/false warning conclusions vs. false negative/no warning conclusions in epidemiological studies of cancer determinants**

Abiding by policy and enforcement and liability potential corollaries of environmental epidemiology, results should endure constant lay public and scientific scrutiny. To such an extent that scholars has called attention to adverse impact of *false positive* results (rather their publishing). An example can be found in a commentary authored by Boffetta et. al. (2008). Appealed by various sources of false positive results and publication bias on cancer causation, the authors propose guidelines aimed at early avoidance of any exaggerated impacts in term of costs and nuisance to stakeholders and the whole community. Other environmental and occupational scholars of authority challenge that judgment: explaining why social nuisance and human toll and costs come indeed as a result of *false negative* studies; and underscoring that publication bias (of untrue negative outputs) remains a hallmark of pharmaceutical product studies <sup>241,242</sup>. Hauptman and Ronckers (2010) call attention on ethical preponderance of reporting an observed, yet controversial association. The precautionary principle, as prescribed by European Environment Agency precisely appeals to uncovering the accumulated public health understanding and health economics impacts of *late lessons from early warnings*.

So has since admonished us humdrum recalls from legion of courageous and independent scientists and regulators that one cannot help prioritizing prime life and quality of existence (e.g. Hill, 1965;. Bross, 1977). Persuading statements like the secular excess of LHs caused by benzene exposure implied a regretful delay of regulation despite longstanding available early warning signs of a true association <sup>245,246</sup>. Requirements for ‘more analysis’ and ‘reaching additional information out’ were relentlessly claimed throughout that regulatory procrastination <sup>247</sup> – the kind of experience whence the editorial team in *Late Lessons from Early Warnings: the precautionary principle 1896-2000* warns us away from << paralysis by analysis >>.

...All the same the technical argument for <<endless analysis>> to marshal, run into the fact that observational studies embody much of an insensitive tool because of an inherent low statistical power.

Have an affordable case-control study, an exposure prevalence of 1:1000 entails an associated Odds Ratio (OR) higher than 10 to attain a conservative chance to be detected ((1-β)>0.60)). Exposure prevalence one logarithm higher still entail an OR of 3 or greater. Those estimations given an alpha of 0.05 and sampling 100:1000 cases to individual controls.





It is not incumbent upon you to complete the work,  
but neither are you at liberty to desist from it

[Chapters of Ethics of the Fathers]



## Corrigendum

### Corrigendum to “Applying a sequential approach in order to detect and interpret a possible increased rate of myeloid malignancies in the Girona province”

[Doctoral Dissertation, Programme in Molecular Biology Biomedicine and Health, 2021].

The author regrets that due to an error in Table 4, an upward veer eventuated in just the starting stretch of the derived Figure 5. The mistake did not engender another changes in statistics, attendant inferences or interpretation shift whatsoever worthy of note.

Table 4 and Figure 5 have now been modified and presented below. Technically, this allows us to restate that the sq curves tolerate an event’s recording error.

The revisited  $\gamma$  estimate for the epidemic span of MPN cases in Girona town (section 5.2.1) is **2.18** instead of 2.02; an iota of risk-fold difference that has no bearing on my interpretation of the results thereof.













An unfortunate oversight lay unnoticed for long following our publishing of the derived article. Not sooner indeed than the eve of preparing the defence of this dissertation. The paper, embedded in the appendix and cited as reference # 107, needs two amendments, precisely in the ending of the Appendix B as follows.

Instead of  $T_2 = a_0 g^4 \frac{(1-g^5)}{1-g}$ , the equation should read  $T_2 = a_0 g^5 \frac{(1-g^5)}{1-g}$

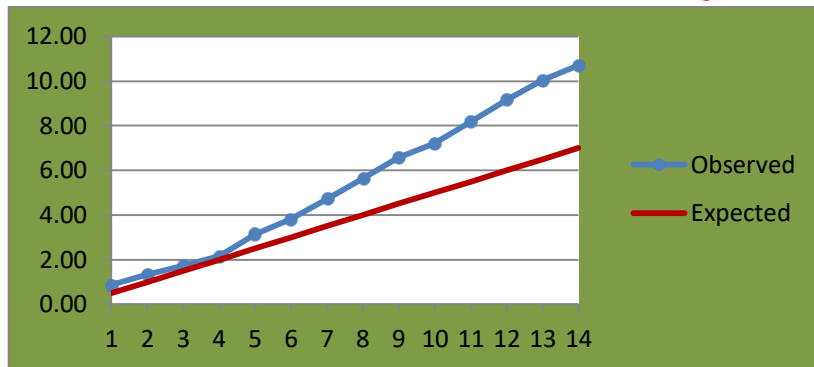
Instead of  $g = \sqrt[4]{\frac{T_1}{T_2}}$ , the equation should be  $g = \sqrt[5]{\frac{T_2}{T_1}}$

The author wishes to apologize for any inconvenience caused.

**Table 4. Sequential assessment. MPN subjects, Sant Hilari de Sacalm.**

Case Number	Age gender	Case date	Interval		
			RI	q	
1	49 f	??/??/1994		<b>0.0007</b>	<b>0.865</b>
				<b>0.2902</b>	
2	63 f	21/02/1997		<b>0.6214</b>	<b>0.465</b>
				<b>0.9108</b>	
3	64 m	??/??/2003		<b>0.0008</b>	<b>0.412</b>
				<b>1.7722</b>	
4	61 f	20/03/2003		<b>0.0008</b>	<b>0.412</b>
				<b>1.7722</b>	
5	71 m	03/04/2003		<b>0.0008</b>	<b>0.994</b>
				<b>0.1116</b>	
6	51 m	17/11/2004		<b>0.2835</b>	<b>0.668</b>
				<b>0.5223</b>	
7	80 f	08/03/2005		0.0993	0.905
8	46 m	08/06/2005		0.0805	0.923
9	42 m	11/08/2005		0.0563	0.945
10	58 m	10/01/2007		0.4568	0.633
11	74 m	19/02/2007		0.0376	0.963
12	77 f	14/03/2007		0.0241	0.976
13	56 f	01/08/2007		0.1319	0.876
14	38 f	13/10/2008		0.4160	0.660

MPN=myeloproliferative neoplasms. RI (Relative Interval)= Expected number of cases during an observed time interval between cases; Ricrit=0.405. q-interval=Null probability that an interval is longer than that observed. If a point measure of RI is unachievable, a median q interval is derived out of longest and shortest possible -in bold- This proxy qi values endow a proxy value for graphic approximation.



**Figure 1 Observed and expected cumulative q-intervals for MPN in Sant Hilari Sacalm (1994-2008).**

MPN=myeloproliferative neoplasms. The q-interval is calculated for each diagnosis as the a-priori probability for a longer time than that observed between consecutive cases. Increasing slope of the curve indicates increased incidence. The curve of the observed data indicates increased incidence from the 5<sup>th</sup> subject to the latest.