epidemiology biostatistics and public health formerly ITALIAN JOURNAL OF PUBLIC HEALTH

EBPH – 2024, Volume 19, Issue 1

https://riviste.unimi.it/index.php/ebph/index

Volume 19

Issue 1

June 2024

Milano University Press

ebph epidemiology biostatistics and public health formerly ITALIAN JOURNAL OF PUBLIC HEALTH

PUBLISHER

Milano University Press
Via Festa del Perdono 7
20122 Milano, Italy
riviste@unimi.it

CONTACT

ebph@unimi.it

Milano University Press

epidemiology biostatistics and public health formerly ITALIAN JOURNAL OF PUBLIC HEALTH

Volume 19, Issue 1 | June 2024

CONTENTS

EDITORIAL

ORIGINAL ARTICLES

SYSTEMATIC REVIEWS AND META- AND POOLED ANALYSES

BIOSTATISTICS

Milano University Press

Cancer Mortality Trends and Predictions for 2024 in Italy

Silvia Mignozzi[(1)], *Claudia Santucci*[(1)], *Eva Negri*[(2)], *Carlo La Vecchia*[(1)]

(1) Department of Clinical Sciences and Community Health, Department of Excellence 2023-2027, University of Milan, Milan, Italy.
(2) Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy.

CORRESPONDING AUTHOR: Claudia Santucci, Department of Clinical Sciences and Community Health, Department of Excellence 2023-2027, University of Milan, "La Statale" Via Celoria 22, 20133 Milan, Italy. Email: claudia.santucci@unimi.it

The prediction of numbers and trends for cancer mortality is useful for evaluating disease burden and reflects the impact of incidence, screening and advancement in early diagnosis and treatments for major cancer sites. Following our recent publication on the estimation of the number of cancer deaths and the corresponding mortality rates for all cancers and selected major cancer sites for the year 2024 in the European Union and the UK [1], here we provide corresponding figures for Italy.

We retrieved official death certifications for various cancer sites from the World Health Organization database in Italy [2]. We calculated sex and age-specific mortality rates for 5-year age groups, ranging from 0-4 years up to 85+ years, and for each calendar year over the period 1970-2019. We derived age-standardized, mortality rates (ASR), using the world standard population, and to analyze ASR trends we fitted joinpoint regression models [3, 4]. To predict 2024 mortality figures, we applied a logarithmic Poisson joinpoint regression model to the number of deaths in each 5-year age group. We estimated age-specific numbers of deaths and their corresponding 95% prediction intervals (PIs), by fitting a linear regression model to the mortality data for each age group, considering the most recent trend segment identified by the joinpoint model. We then calculated both age-specific and age-standardized death rates, along with their related 95% PIs, using the predicted age-specific number of death counts. Population predictions were obtained from the Eurostat database [5]. In addition, the number of deaths averted for all cancers from the peak observed in 1988 to 2024 was estimated for all cancers combined.

Statistical analyses were performed using the software R version 4.3.2 (R Development Core Team, 2022) and Joinpoint Regression Program version 5.1.0 (Statistical Methodology and Applications Branch, Surveillance Research Program, National Cancer Institute).

Table 1 gives the predicted cancer deaths and rates, along with their corresponding 95% PIs for 2024, in comparison with observed figures for 2019 in Italy. For 2024, we estimated 98,700 cancer deaths in men, with corresponding ASR of 103.9/100,000 (-9.0% vs 2019), and 82,000 in women, ASR of 72.1/100,000 (-4.4% vs 2019). Overall, the predicted ASRs are favourable for all cancer sites and both sexes, except for pancreatic cancer among men (+2.7% vs 2019) and women (+1.7% vs 2019), and for lung cancer among women (+5.9% vs 2019).

Figure 1 shows the trends in cancer mortality rates over calendar periods, among men and women, from 1970-74 to 2015-19, along with the predicted ASRs for 2024 with the corresponding PIs as well as the total avoided cancer deaths for men and women between the top rate in 1988 and 2024.

Among men, stomach cancer ASR decreased over the whole period, while most other cancer sites began falling in the early 1990s. Among women, the ASRs for stomach, leukemia, and uterine cancers have exhibited a decline since 1970, while rates for breast, colorectal, and ovarian cancers started to decline during the 1990-94 quinquennium. Pancreas and lung cancer showed unfavorable trends over the whole period. Since 1988, about 1,248,100 cancer deaths have been avoided in Italy, 921,900 in men and 326,200 in women, respectively.

The 2024 predicted mortality cancer figures are favourable in Italy. Rates fell by 9.0% in men and by 4.3% in women. Pancreatic cancer is the only cancer site that has shown a lack of progress for both sexes. In Italy in 2022, pancreatic cancer ranked sixth in the number of new cases (15,710) and fourth in the number of deaths (14,903) [6], in line with the European and USA's estimates [7, 8]. Patterns of smoking prevalence, which is the main risk factor for pancreatic cancer, along with overweight, obesity, diabetes and heavy alcohol consumption can only partly explain the observed pattern. Pancreatic cancer survival is still low, with limited progress in early

DOI: 10.54103/2282-0930/23645
Accepted: 14th June 2024

Table 1. Number of predicted deaths and mortality rate for the year 2024 and comparison figures for 2019 for Italy, with 95% prediction intervals.

Sex	Cancer	Observed number of deaths 2019	Predicted number of deaths 2024 (95% PI)	Observed ASR 2019	Predicted ASR 2024 (95% PI)	% Difference 2024 vs 2019
Men	Stomach	5,277	4803 (4567-5039)	5.99	4.80 (4.43-5.17)	-19.80
	Colorectum	12,040	12425 (12055-12795)	13.32	12.76 (12.29-13.22)	-4.22
	Pancreas	6,267	6799 (6618-6981)	7.71	7.92 (7.66-8.18)	2.73
	Lung	22,853	21356 (20706-22006)	26.73	22.75 (21.89-23.61)	-14.89
	Prostate	7,694	7938 (7657-8219)	6.69	6.39 (6.15-6.63)	-4.41
	Bladder	4,748	5060 (4749-5372)	4.55	4.52 (4.18-4.85)	-0.77
	Leukemias	3,627	3758 (3617-3900)	4.35	3.88 (3.60-4.16)	-10.77
	All cancers	99,380	98684 (96906-100463)	114.18	103.91 (101.19-106.63)	-9.00
Women	Stomach	3,708	3405 (3217-3594)	3.08	2.63 (2.40-2.86)	-14.52
	Colorectum	9,971	9791 (9468-10115)	8.06	7.46 (7.19-7.73)	-7.45
	Pancreas	6,550	7045 (6837-7254)	5.78	5.88 (5.69-6.07)	1.71
	Lung	10,162	11414 (11068-11760)	10.60	11.22 (10.81-11.64)	5.92
	Breast	12,830	13501 (13118-13884)	13.97	14.05 (13.50-14.6)	0.56
	Uterus	3,098	3235 (3091-3380)	3.50	3.61 (3.43-3.79)	3.09
	Ovary	3,410	3448 (3283-3613)	4.05	3.90 (3.68-4.13)	-3.62
	Bladder	1,341	1490 (1389-1591)	0.97	1.03 (0.95-1.11)	6.39
	Leukemias	2,721	2793 (2671-2915)	2.45	2.30 (2.07-2.52)	-6.17
	All cancers	79,918	82008 (80713-83302)	75.38	72.09 (70.48-73.71)	-4.36

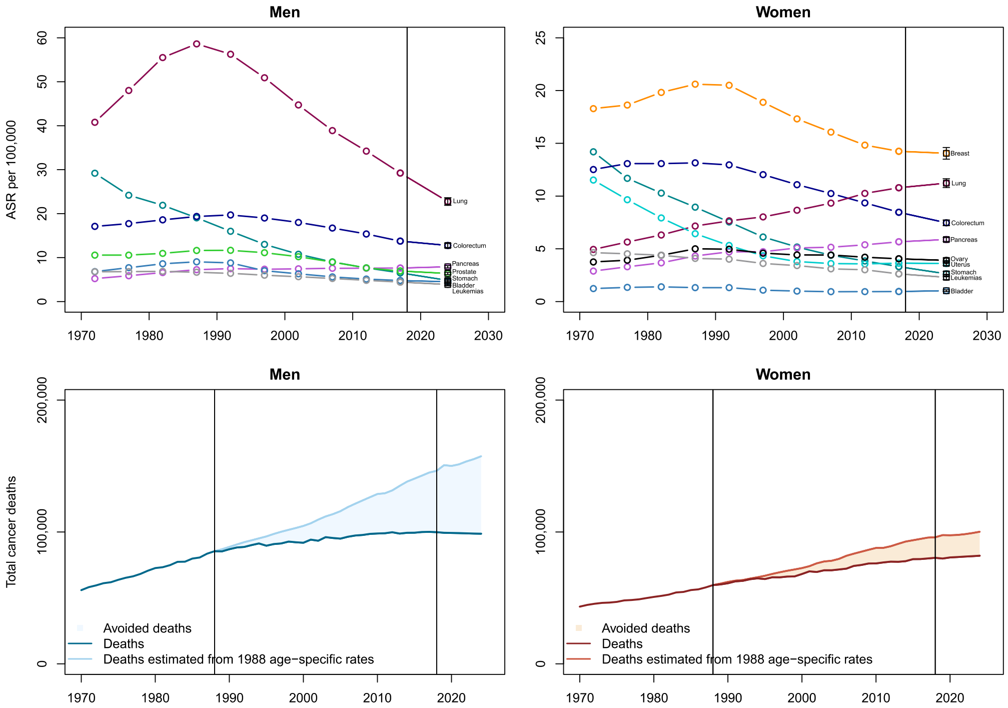
ASR, age-standardized rate; PI, prediction interval

diagnosis and treatment although the identification and targeting of specific tumour-associated antigens and mutations have been improved [9, 10].

Lung cancer among women also continues to show unfavorable trends, linked to the trends and prevalence of smoking among them. In Italy in 2022, lung cancer ranked third in the number of new cases (43,808) and first in the number of deaths (35,668) [6]. Sex-difference in mortality trends is also in line with the fact that the peak in smoking-attributable mortality among men was reached in the late 1980s in Western Europe, whereas for women it is now being reached [11].

In conclusion, projected cancer mortality rates for 2024 remain favorable in Italy, especially among men, due to smoking cessation. Lifestyle factors such as overweight, obesity, diabetes, and alcohol consumption may have contributed to the unfavorable trends in pancreatic cancer, whereas rising lung cancer trends among women likely reflect smoking patterns.

Figure 1. Age-standardized cancer mortality rate (ASR) trends from 1970-74 to 2015-19 and predicted rates for 2024 with 95% prediction intervals, for major cancer sites in men and women (top panels) and total avoided cancer deaths for men and women between the top rate in 1988 and 2024 in Italy (bottom panels).



REFERENCES

1. Santucci C, Mignozzi S, Malvezzi M, Boffetta P,Collatuzzo G, Levi F, et al. European cancer mortality predictions for the year 2024 with focus on colorectal cancer. Ann Oncol. 2024;35(3):308-16.
2. World Health Organization Statistical Information System. WHO mortality database. Geneva: World Health Organization. Available at: https://www.who.int/data/data-collection-tools/who-mortality-database (Last accessed: December 2023).
3. Kim HJ, Chen HS, Byrne J, Wheeler B, Feuer EJ. Twenty years since Joinpoint 1.0: Two major enhancements, their justification, and impact. Stat Med. 2022;41(16):3102-30.
4. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med. 2000;19(3):335-51.
5. European Commission. EUROSTAT population database. Available at: https://ec.europa.eu/eurostat/web/main/data/database (Last Accessed: January 2024).
6. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2024). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today (Last accessed May 2024).
7. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63.
8. Cronin KA, Scott S, Firth AU, Sung H, Henley SJ, Sherman RL, et al. Annual report to the nation on the status of cancer, part 1: National cancer statistics. Cancer. 2022;128(24):4251-84.
9. Nevala-Plagemann C, Hidalgo M, Garrido-Laguna I. From state-of-the-art treatments to novel therapies for advanced-stage pancreatic cancer. Nat Rev Clin Oncol. 2020;17(2):108-23.
10. Warren EAK, Lesinski GB, Maithel SK. Top advances of the year: Pancreatic cancer. Cancer. 2023;129(24):3843-51.
11. Janssen F, El Gewily S, Bardoutsos A. Smoking epidemic in Europe in the 21st century. Tob Control. 2021;30(5):523-9.

Short Term Regional and Age-Specific Disparities in Suicide Epidemiology in Poland

*Natalia Olszańska**[(1)], *Przemysław Waszak**[(2)], *Paweł Zagożdżon*[(2)]

(1) Faculty of Medicine, Medical University of Gdansk, Poland
(2) Department of Hygiene and Epidemiology, Medical University of Gdansk, Poland
** Joint first co-authorship (these authors contributed equally to this work).*

CORRESPONDING AUTHOR: Przemysław Waszak, Division of Hygiene and Epidemiology, ul. Dębinki 7, Gdańsk, Poland. Email: p.waszak@gumed.edu.pl

SUMMARY

Introduction: Despite declining trends in the first two decades of the 21st century, Poland remains a country with relatively high suicide rates. Developing national suicide prevention programmes starts from analysing trends in suicide rates and identifying high risk groups. The aim of this study was to examine suicide epidemiology trends in Poland with a specific focus on age groups and regional differences.
Method: This epidemiological analysis examined suicide statistics from 2017 to 2022. We calculated and analysed standardised suicide rates (SDR) across different age groups and regions in Poland using data acquired from Police Headquarters statistics. Percentage changes for the whole study period were determined. Official data on the Polish population was obtained from the Central Statistical Office.
Results: Throughout the analysed period, SDR calculated for all ages remained stable, declining by only 2% from 2017 to 2022. The highest SDR were noted in the 55-59 and 60-64 age groups (19,4 and 19,1 per 100 000, respectively, in 2022). Between age groups, notable disparities in trends of changes of SDR values were observed. The greatest increases of 21.6% and 19.6% were noted in the youngest (13-18) and eldest (85+) age groups, respectively. The largest regional increase by 14.4% concerned the Warmian-Masurian region, followed by the Opolskie region by 13.51%. Both regions have some of the lowest GDP values among Polish regions.
Conclusion: In Poland suicide rates have increased significantly among adolescents, the eldest and those living in economically disadvantaged regions. The obtained results highlight the need for implementing tailored preventative programmes.

Keywords: Suicide trends; High-risk age groups; Regional variations; Poland; Prevention programs.

INTRODUCTION

Suicide is a major public health concern worldwide. According to the World Health Organisation (WHO), nearly 800,000 people die by suicide each year, and it is the second leading cause of death among 15-29-year olds globally. Suicide rates vary across countries and regions, with the highest rates generally observed in low- and middle-income countries. Risk factors for suicide include mental illness, substance abuse, social isolation, and access to lethal means. Prevention efforts include improving access to mental health care, reducing stigma surrounding mental illness, and implementing policies to restrict access to means of suicide.

DOI: 10.54103/2282-0930/22856
Accepted: 10th May 2024

Despite declining trends in suicide rates in the first two decades of the 21st century, Poland continues to have a high suicide death rate [1]. According to data from the World Health Organisation, the suicide rate in Poland was around 11.3 per 100,000 population in 2019. This is higher than the average suicide rate in Europe, which is around 10.5 per 100,000 population [2]. Among Poles aged 30-34 suicide is the leading cause of death and among adolescents, suicide is the second leading cause of death, preceded only by road injuries [3]. In Poland, the suicide mortality pattern by age groups resembles that of less developed countries where suicides are more prevalent among individuals of working age rather than late old age [4]. There is also a significant disparity in the standardised death rate between male and females, with males experiencing a rate seven times higher than that of females, whilst globally it is approximately 2.3 times higher. Roughly 90% of male suicides and 80% of female suicides are carried out by the means of hanging [1,5]. This study aims to explore changes in suicide epidemiology during the 2017-2022 period in Poland, with a specific focus on age and regional disparities. By the analysis of recent suicide trends, we hope to distinguish high risk populations and therefore aid national health authorities in the development of tailored suicide prevention programmes.

METHODS

Figures on suicidal behaviour reported by the Polish Police were obtained for the study. These data are publicly available through the National Police Headquarters website (http://bip.kgp.policja.gov.pl/). The figures for suicide deaths and suicidal behaviour (deaths + attempts) are presented separately and sorted by region, age group, etc.

Poland consists of 16 administrative regions (provinces). Suicide data is collected by 17 regional police headquarters. Warsaw, the capital, has its own separate police headquarter, thus Warsaw statistics were included in this study in the Mazovia province (to which Warsaw administratively belongs).

Since 2017, the Polish Police has significantly changed its data collection methodology, so the period 2017-2022 was selected for analysis. For the purposes of this study, we used data only on suicide deaths by region and age group. For clarity of presentation of results, only those aged 13+ were included. Only isolated cases of suicide were recorded in those younger than 13.

Data on the numbers of Poland's population by age group and region were obtained from the Central Statistical Office. Through the website (https://bdl.stat.gov.pl/bdl/start), anyone can get a glimpse of the official demographic data.

In our study, we employed an age-specific standardisation method to analyse suicide mortality trends across different regions and age groups within Poland. Specifically, the number of suicides recorded in each age group was standardised against the population size of the same age group within the respective region or the entire country, as provided by the Central Statistical Office of Poland (GUS). This approach involved calculating age-specific suicide rates by dividing the total number of suicides in each age group by the total population of that age group in the same region. Suicide death rates (SDR) were reported as standardised numbers per 100,000 people in our study (age-standardised suicide rates). Epidemiological standardisation indicators are measures that are used to adjust for differences in population characteristics when comparing health outcomes across different groups or time periods. The standardisation process involves applying a mathematical formula to adjust for the differences in the distribution of the standardisation indicator (e.g. age, or region) between populations or time periods. This adjustment allows for more accurate comparisons of health outcomes and disease rates, and can help identify differences in health disparities across populations. This standardisation is consistent with the official reporting of suicide by the World Health Organisation.

We utilized a linear regression model to define the trend line, which is mathematically represented by the equation $y=mx+b$. This model was chosen to identify and illustrate linear trends over time in the suicide data across different age groups. The m is the slope of the line and b is the intercept. The x and y represent the distance of the line from the x-axis and y-axis, respectively.

Finally, the study presents the SDR for each age group and region of Poland from 2017 to 2022 and calculates percentage changes in the SDR.

RESULTS

Across the study period, an overall decrease in suicide death rates by 1.7% was observed in the studied population (from 15.83 per 100 000 in 2017 to 15.55 per 100 000 in 2022). The greatest suicide mortality rates were seen in the 55-59 and 60-64 age groups (19.38 and 19.07 per 100 000 in 2022, respectively) while the 13-18 age group had the lowest (6.41 per 100 000 in 2022).

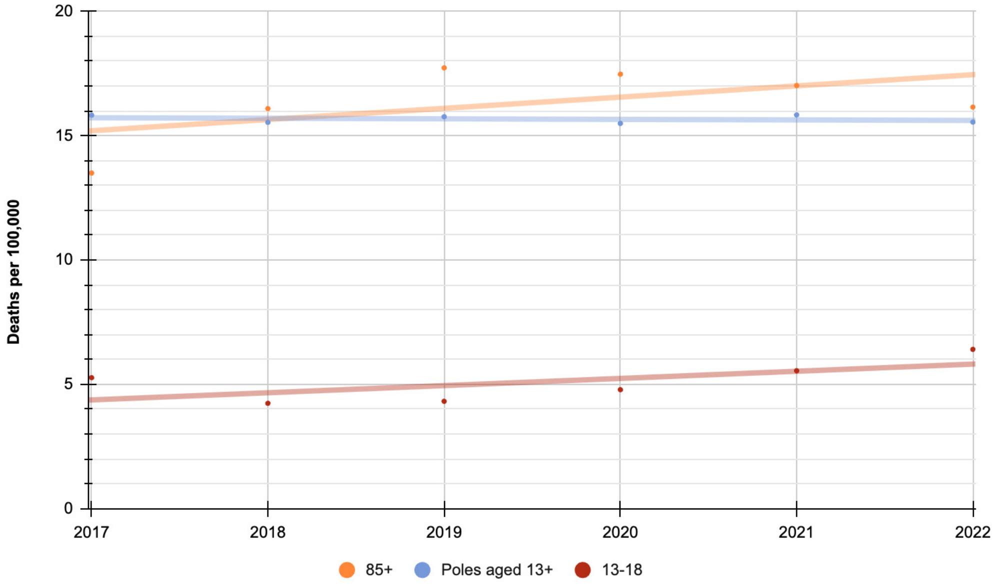
Rates for the 55-59 age group were consistently high; however, a significant fall from 23.04 per 100 000 in 2017 to 19.38 per 100 000 in 2022 was observed. In this age group suicide mortality fell regularly between 2017 and 2020, then rose briefly in 2021 before dropping again in 2022. This overall decrease by 15.9% was the greatest among all ages. Declining rates were also observed among Poles aged 40 or older excluding the 75-79 and 85+ population.

The largest increases in suicide mortality rates concerned the youngest and eldest age groups.

Table 1: Suicide death rates by age with percentage change across 2017-2022

Age group	SDR 2017	SDR 2018	SDR 2019	SDR 2020	SDR 2021	SDR 2022	Change (%)
13-18	5,27	4,24	4,32	4,79	5,55	6,41	21,6%
19-24	13,59	13,80	14,97	14,94	15,42	14,60	7,4%
25-29	15,13	14,78	16,00	17,27	16,89	16,35	8,1%
30-34	15,91	14,65	16,26	16,03	17,09	17,91	12,6%
35-39	14,74	13,89	16,10	14,38	15,84	16,43	11,4%
40-44	16,08	15,87	14,94	15,93	15,96	14,90	-7,4%
45-49	16,66	17,42	17,43	15,97	16,69	15,54	-6,7%
50-54	19,13	18,73	17,25	17,48	17,76	18,02	-5,8%
55-59	23,04	22,00	18,77	18,46	20,76	19,38	-15,9%
60-64	19,33	20,35	19,33	19,55	17,77	19,07	-1,4%
65-69	16,10	14,80	16,37	15,45	15,15	14,44	-10,3%
70-74	14,81	16,35	15,96	14,14	14,20	12,76	-13,8%
75-79	13,08	14,75	14,45	14,21	13,15	14,75	12,8%
80-84	16,40	12,44	15,12	15,83	16,40	15,97	-2,6%
85+	13,50	16,09	17,73	17,48	17,02	16,15	19,6%
Poles aged 13+	15,83	15,54	15,76	15,49	15,84	15,55	-1,7%

Figure 1: Overall downward trend in the studied population and upward trends in the 13-18 and 85+ age groups (linear trend line)



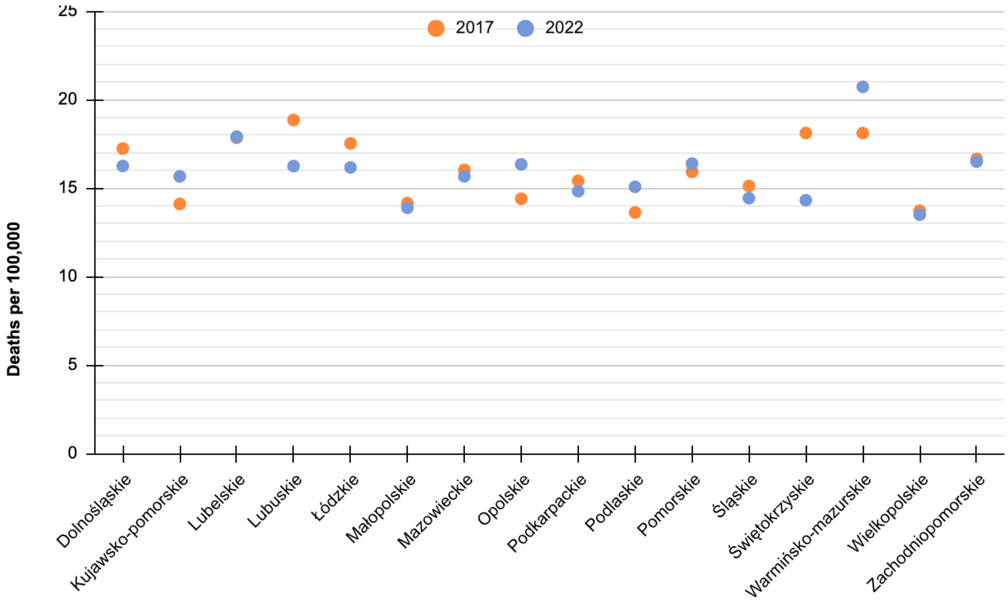
In the 13-18 population, rates increased by 21.6% (from 5.27 per 100 000 in 2017 to 6.41 per 100 000 in 2022). Initially, rates dropped between 2017 and 2018, but since 2018 they have been steadily rising. The second largest increase of 19.6% was seen in the 85+ age group (rising from 13.5 per 100 000 in 2017 to 16.15 per 100 000 in 2022).

This study also analysed differences in suicide rates between voivodeships in Poland. As seen in Table 2, the highest suicide mortality rate was observed in the Warmian-Masurian voivodeship, peaking at 20.72 per 100 000 during the study period. Whereas, the lowest suicide rates were seen in the Wielkopolska voivodeship, being 13.50 per 100 000 in 2022.

Table 2: Suicide mortality rates across voivodeships during 2017-2022 with percentage change

Vovoideship	SDR 2017	SDR 2018	SDR 2019	SDR 2020	SDR 2021	SDR 2022	Change (%)
Dolnośląskie	17,23	16,83	17,82	16,68	16,20	16,25	-5,70%
Kujawsko-pomorskie	14,11	15,53	14,33	13,99	15,88	15,66	11,04%
Lubelskie	17,85	17,38	16,52	16,82	18,09	17,91	0,32%
Lubuskie	18,85	17,41	19,95	16,79	20,67	16,24	-13,83%
Łódzkie	17,53	18,12	16,06	15,13	15,51	16,17	-7,76%
Małopolskie	14,15	13,55	14,72	16,19	14,31	13,89	-1,88%
Mazowieckie	16,03	16,58	16,10	16,15	16,19	15,67	-2,24%
Opolskie	14,40	14,12	14,87	12,29	14,38	16,34	13,51%
Podkarpackie	15,41	15,20	12,72	14,35	14,93	14,83	-3,80%
Podlaskie	13,62	13,48	16,63	16,60	14,75	15,07	10,64%
Pomorskie	15,92	16,53	15,66	14,98	15,87	16,39	2,92%
Śląskie	15,12	15,04	13,90	14,35	15,24	14,43	-4,58%
Świętokrzyskie	18,12	16,56	17,48	17,22	18,01	14,32	-20,97%
Warmińsko-mazurskie	18,11	17,28	19,53	19,65	18,27	20,72	14,42%
Wielkopolskie	13,73	12,92	14,60	13,53	13,77	13,50	-1,65%
Zachodniopomorskie	16,65	17,64	17,94	16,43	16,96	16,50	-0,93%

Figure 2: Changes in suicide death rates across voivodeships between 2017 and 2022.



With regard to changes in suicide death rates across voivodeships in the studied years (Figure 2), increases were observed in the Kujawsko-pomorskie, Lubelskie, Opolskie, Podlaskie, Pomorskie, and Warmińsko-mazurskie regions. The remaining voivodeships encountered overall decreases, with the largest observed in the Świętokrzyskie region (20.97%) and Lubuskie region (13.83%).

The Warmińsko-mazurskie voivodeship saw the largest increase of 14.42% (from a rate of 18.11 per 100 000 in 2017 to 20.72 per 100 000), but year-to-year changes experienced alternating increases and decreases.

The second greatest increase of 13.51 percent was observed in the Opolskie voivodeship (from 14.4 to 16.3 per 100 000 population).

Across the study period, the incidence of suicide death was vastly higher among males than females. In 2017, 85.74% of all suicide deaths concerned males. A difference in trends between the two genders is noted. The suicide death rate in males decreased by 4.03 % from 24.33 per 100 000 in 2017 to 23.35 per 100 000 in 2022. Whereas the suicide death rates in females increased by 14.51% from 3.79 per 100 000 in 2017 to 4.34 per 100 000 in 2022.

The dominant suicide method was hanging, accounting for 79.78% of all suicides across the studied period. The second most frequent method was jumping from a height (6.89%) and the third most frequent throwing under a moving vehicle (2.52%). Twenty percent of cases were previously treated for a psychiatric disorder and 18.42% had a history of alcohol abuse.

The most common reason of suicide was mental illness/ mental disorder (19.73%) followed by family disagreements/ family violence (4.73).

The state of consciousness in most instances was not determined, however out of all detected substances, alcohol was most prevalent (10.28%).

Figure 3: Suicide death rates by gender across 2017 to 2022.

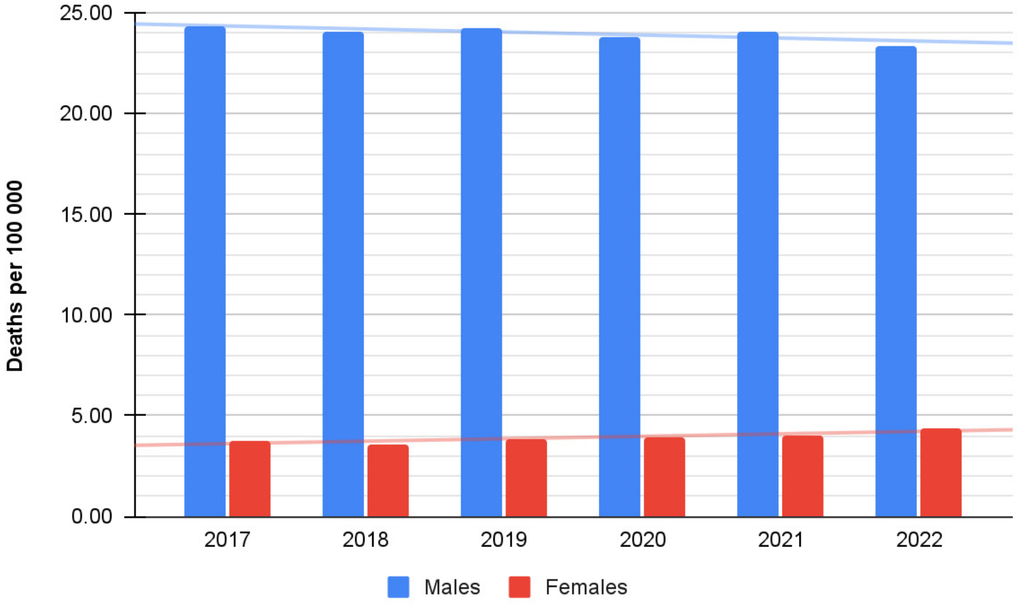


Table 3: Sociogeographic characteristics of population who died by suicide in Poland in the period 2017-2022 (based on data from Police Headquarters).

	Number of suicides TOTAL = 65190	%
Health status		
No data available	33626	53,91
Physical illness	3894	6,24
Treated for psychiatric disorder	12628	20,25
Treated for alcohol addiction	2082	3,34
Treated for drug addiction	258	0,41
Abused alcohol	11492	18,42
Permanently disabled	482	0,77
Used illicit drugs	728	1,17
Contact with institutions		
Unable to determine	52876	84,77
Contact with other institution	450	0,72

	Number of suicides TOTAL = 65190	%
Contact with an church institution	52	0,08
Contact with a crisis intervention center	20	0,03
Contact with a social welfare center	422	0,68
Contact with a medical facility	7444	11,93
Contact with the police	1440	2,31
Reason for suicide		
Physical illness	2034	3,26
Mental illness/mental disorder	12304	19,73
Committing a felony or misdemeanor	316	0,51
Other	2666	4,27
Conflict with people outside the family	144	0,23
Bullying, cyberbullying, abuse	10	0,02
Sudden loss of livelihood	486	0,78
Family disagreements/family violence	2952	4,73
Unwanted pregnancy	6	0,01
Undetermined	35712	57,25
HIV carrier/AIDS patient	14	0,02
Deterioration or sudden loss of health	1712	2,74
Problems at school or work	440	0,71
Death of a loved one	1006	1,61
Permanent disability	200	0,32
Threat or loss of residence	90	0,14
Love disappointment	2582	4,14
Poor economic conditions/debts	1916	3,07
Work or school status		
Unemployed	10316	16,54
No data	32968	52,86
Short-term job	4434	7,11
Permanent job	8786	14,09
Self-employed	1734	2,78
Farmer	2262	3,63
University student	468	0,75
Primary school student	1406	2,25
Methods		
Other	714	1,14
Hanging	49760	79,78
Throwing under a moving vehicle	1574	2,52
Jump from a height	4300	6,89
Self-harm/ superficial injury	470	0,75
Self-immolation	152	0,24

	Number of suicides TOTAL = 65190	%
Suffocation	550	0,88
Injury to the circulatory system	1072	1,72
Drowning	648	1,04
Using a firearm	974	1,56
Gas/fumes poisoning	422	0,68
Poisoning by chemical agents/toxins	272	0,44
Poisoning by illicit drugs	30	0,05
Ingestion of other drugs	780	1,25
Ingestion of sleeping pills/psychotropic drugs	656	1,05
Marital status		
No data available	6762	10,84
Single	20042	32,13
Informal relationship	2398	3,84
Divorced	5370	8,61
Separated	286	0,46
Widowed	4298	6,89
Married	23218	37,22
State of consciousness		
No data available	51208	82,10
Under the influence of alcohol	6410	10,28
Under the influence of medications	742	1,19
Under the influence of illicit drugs	154	0,25
Sober	4092	6,56
Education		
No data available	45368	72,74
Middle school	790	1,27
Primary	3886	6,23
Partial primary	324	0,52
Post-secondary	44	0,07
Secondary	4580	7,34
Higher education	1690	2,71
Vocational	5692	9,13
Source of income		
No data available	25890	41,51
Retirement	8320	13,34
Dependent on another person	4754	7,62
Not on a fixed income	5532	8,87
Work	14040	22,51
Pension	3148	5,05
Allowance/alimony	690	1,11

Sociodemographic characteristics such as marital status, education, source of income, and work or school status are also provided in Table 3, likewise in many cases no data was available.

DISCUSSION

Our study found an overall decrease during the period spanning from 2017 to 2022. This outcome is consistent with other research that also revealed a decline in suicide death rates in previous years in Poland [4]. Partially, this phenomenon can be attributed to the improving economic conditions in Poland, including the declining unemployment rate, rising average salaries, and decreasing rates of both relative and extreme poverty [1].

Among all age groups, the greatest increase of 21.6% was observed among youth, with suicide rates consistently increasing since 2018. This rising trend represents a shift from past years, when the number of suicide deaths in this demographic was falling [4]. Over the last couple of years, a decline in children's mental health has been perceived, with studies showing increased psychological distress, suicide-related behaviours, and suicide attempts [6-8].

The increase in suicide deaths clearly occurred after the COVID-19 pandemic. In Poland, the lockdown policy from the very beginning was based on the closure of schools and the shift to on-line learning. There was also data showing an increase in domestic violence during the lockdown period in Poland [9]. In addition, for minors in the early days of the pandemic, it was forbidden to leave home unaccompanied by an adult. According to the report by the Foundation Dajemy Dzieciom Siłę, almost one in three respondents (30.8%) felt that their well-being had worsened during the period under review [10]. Girls complained of worse well-being significantly more often than boys. During the first period of the pandemic, 4.4% of respondents mutilated themselves more often than before the pandemic, while 2.9% of respondents aged 15-17 attempted suicide [10].

According to the Supreme Audit Office, the psychiatric health care system for children and adolescents in Poland does not provide comprehensive and accessible care, and therefore requires systemic changes [11]. In addition, more accessible and tailored prevention programmes should also be implemented, such as school-based suicide-prevention [12].

In addition, an alarming increase of 19.6% was observed among Poles aged 85 years or older (from a rate of 13.50 per 100 000 to 16.15 per 100 000). It is concerning that the number of suicide deaths among the eldest is rising. Continuation of this trend may result in this age group having the highest suicide mortality rate among all demographics which is already seen in several European nations [13].

There are distinct variations in the characteristics of suicide among the elderly population. Among older adults, the prevalence of any psychiatric disorder or psychiatric therapy declines with age [14]. People aged 85 and above are less likely to have had a previous suicide attempt, had a past psychiatric admission, or used psychiatric services a month prior to dying by suicide [15]. However, a major percentage had contact with their general practitioners a month before death, emphasising the importance of suicide prevention programmes in primary care [15,16].

Suicidal behaviour in older adults has been strongly associated with functional impairment, physical illness, a feeling of loneliness, and a loss of meaning of life. Older adults with a wish to die more often have negative judgments about their age, in particular relating to a sense of worth and dignity [17].

Poland, like other European nations, is experiencing an aging population. It is likely that the absolute number of suicides among the eldest will rise. Therefore, it is essential to prioritise mental health support systems tailored to the needs of older adults. This includes strengthening social welfare programmes, enhancing mental health screenings, and recognising the impact of physical health and disabilities on suicide risk in older individuals.

As for regional disparities, the largest percentage increase in suicide rates concerned the warminsko-mazurskie voivodeships (14.42%). Differences between regional suicide rates may be associated with socioeconomic inequalities [18,19]. Despite improving economic conditions in Poland, the warminsko-mazurskie voivodeship remains one of the most economically disadvantaged regions in the country. It is characterised by the highest unemployment rate, the highest rate of people receiving social assistance benefits, and one of the lowest GDP per capita [20].

Generally, in many countries, a correlation between lower socio-economic status and higher suicide rates has been found. Different patterns of suicides were observed within various social patterns, with higher frequencies among individuals who are not married, those who are unemployed, and those belonging to lower socioeconomic strata [18]. In almost all studies, it was observed that people from lower socioeconomic backgrounds had a greater likelihood of dying by suicide compared to their counterparts in higher socioeconomic groups [18].

The landmark study done by Fiete Näher et al. examined these relationships in detail [21]. The study found that when unemployment in an area goes up by 1%, the suicide rate in that area increases by 1.20%. On the other hand, when incomes increase by 1%, the suicide rate decreases by 0.39%. Conversely, a 1% decrease in the incomes of single individuals is linked to a 0.54% increase in the suicide rate [21].

In a longitudinal study conducted in the United States, researchers discovered that both an individual's socioeconomic status and their subjective social status were predictive factors for heightened levels of depressive symptoms and an increased risk of suicidality [22]. This casualty has also been found to work in the reverse direction, with depressive disorders frequently exhibiting considerable impairments in social functioning [23].

Limitations of the Study

The analysis covered only the years 2017-2022, which might not account for long-term trends or the impact of prolonged risk factors. In the study, we have focused on describing general trends, but we did not conduct a deeper analysis of psychological, social, or economic factors related to the increase in suicide rates. The inclusion of possible further factors such as gender, previous suicide attempts, etc. could significantly advance the scientific field being studied. Thus the study does not provide a deeper understanding of the reasons behind the increase in suicide rates in specific groups and regions. In an ecological study analysing only general trends, it is not possible to analyse individual participants. Only an approach that incorporates such a large-scale analysis (in line with the hierarchy of evidence in evidence-based medicine) would provide insight into the real causes of the observed phenomena. However, a better understanding of short-term regional trends and by age group is an important first step towards formulating better, targeted suicide prevention programmes in Poland. To our knowledge, no such studies on Poland have been undertaken in the literature to date.

CONCLUSIONS

In Poland suicide rates have increased significantly among adolescents, the eldest and those living in economically disadvantaged regions. The obtained results highlight the need for implementing tailored preventative programmes.

FUNDING

This study was financially supported by the Erasmus+ project "Students' of Higher Education critical digital Literacy Development against Disinformation." Project Reference: 2022-1-FR01-KA220-HED-000088804.

DISCLOSURE STATEMENT

The authors report there are no competing interests to declare.

DATA AVAILABILITY STATEMENT

The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

REFERENCES

1. Ashworth E, Thompson J, York S, Henderson K, Jalota M, Shelton J, et al. Assessing the social validity of a multi-modal school-based suicide prevention intervention: A scoping study Executive Summary2022.
2. Cheung G, Merry S, Sundram F. Do suicide characteristics differ by age in older people? Int Psychogeriatr. 2018;30(3):323-30.
3. Dostępność lecznictwa psychiatrycznego dla dzieci i młodzieży (w latach 2017–2019). (n.d.). Retrieved 5 November 2023, from https://www.nik.gov.pl/plik/id,22730,vp,25429.pdf
4. Dzwonnik K, Sowulewski O, Dettlaff-Dunowska M, Waszak PM, Szlagatys-Sidorkiewicz A, Plata-Nazar K. COVID-19 lockdown and domestic violence in Poland – an analysis of crisis helpline and Google data. Pediatria Polska - Polish Journal of Paediatrics. 2022;97(2):111-7.
5. Gawlinski A, Soltyszewski I, Wiergowski M. Epidemiology of suicides in Poland in 1990-2018 - changes and new trends. Arch Med Sadowej Kryminol. 2020;70(4):222-34.
6. Gleeson H, Roesch C, Hafford-Letchfield T, Ellmers T. Assessing suicide ideation among older adults: a systematic review of screening and measurement tools. Int Psychogeriatr. 2022;34(5):439-52.
7. Global Health Estimates 2019: Deaths by cause, age, sex, by country and by region, 2000–2019. (2019).
8. Hill RM, Rufino K, Kurian S, Saxena J, Saxena K, Williams L. Suicide Ideation and Attempts in a Pediatric Emergency Department Before and During COVID-19. Pediatrics. 2021;147(3).
9. Hirschfeld RM, Montgomery SA, Keller MB, Kasper S, Schatzberg AF, Moller HJ, et al. Social functioning in depression: a review. J Clin Psychiatry. 2000;61(4):268-75.
10. Hofer P, Rockett I, Varnik P, Etzerdorfer E, Kapusta N. Forty years of increasing suicide mortality in Poland: Undercounting amidst a hanging epidemic? BMC Public Health. 2012;12.
11. Koo YW, Kolves K, De Leo D. Suicide in older adults: differences between the young-old, middle-old, and oldest old. Int Psychogeriatr. 2017;29(8):1297-306.
12. Lorant V, Kapadia D, Perelman J. Socioeconomic disparities in suicide: Causation or confounding? PLOS ONE. 2021;16:e0243895.
13. Lorant V, Kunst AE, Huisman M, Costa G, Mackenbach J, Health EUWGoS-EIi. Socio-economic inequalities in suicide: a European comparative study. Br J Psychiatry. 2005;187:49-54.
14. Borges D, Rasella D, Santos D. Impact of Income Inequality and Other Social Determinants on Suicide

Rate in Brazil. PLOS ONE. 2015;10:e0124934.

15. Madigan A, Daly M. Socioeconomic status and depressive symptoms and suicidality: The role of subjective social status. Journal of Affective Disorders. 2023;326.
16. Makaruk, K., Włodarczyk, J., & Szredzińska, R. (2020). Negatywne Doświadczenia Młodzieży W Trakcie Pandemii. In Fundacja Dajemy Dzieciom Siłę.
17. Naher AF, Rummel-Kluge C, Hegerl U. Associations of Suicide Rates With Socioeconomic Status and Social Isolation: Findings From Longitudinal Register and Census Data. Front Psychiatry. 2019;10:898.
18. Panchal U, salazar de pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. European Child & Adolescent Psychiatry. 2021;32.
19. Pikala M, Burzynska M. The Burden of Suicide Mortality in Poland: A 20-Year Register-Based Study (2000-2019). Int J Public Health. 2023;68:1605621.
20. Raport o sytuacji społeczno-gospodarczej województwa warmińsko-mazurskiego 2022. (n.d.). Retrieved 5 November 2023, from https://olsztyn.stat.gov.pl/publikacje-i-foldery/warunki-zycia/raport-o-sytuacji-spoleczno-gospodarczej-wojewodztwa-warminsko-mazurskiego-2022,2,11.html?contrast=default
21. Reynolds K, Pietrzak R, El-Gabalawy R, Mackenzie C, Sareen J. Prevalence of psychiatric disorders in U.S. older adults: Findings from a nationally representative survey. World psychiatry: official journal of the World Psychiatric Association (WPA). 2015;14:74-81.
22. Schmidtke A, Sell R, Löhr C. Epidemiologie von Suizidalit??t im Alter. Zeitschrift für Gerontologie und Geriatrie. 2008;41:3-13.
23. World Health Organization. Suicide Worldwide in 2019. (2021). https://www.who.int/publications/i/item/9789240026643
24. Yard E, Radhakrishnan L, Ballesteros M, Sheppard M, Gates A, Stein Z, et al. Emergency Department Visits for Suspected Suicide Attempts Among Persons Aged 12–25 Years Before and During the COVID-19 Pandemic — United States, January 2019–May 2021. MMWR Morbidity and Mortality Weekly Report. 2021;70.

Overview of Trauma Injuries Caused by Traffic Accidents in Baixada Santista, Brazil

Luís Fernando Rosati Rocha[1], *Ana Paula de Carvalho Miranda Rosati Rocha*[1], *Ana Paula Taboada Sobral*[1],[2], *Juliana Maria Altavista Sagretti Gallo*[2],[3], *Nathálie Beatriz do Carmo Silva*[2], *Sandra Kalil Bussadori*[2,4], *Ana Luiza Cabrera Martimbianco*[1], *Gustavo Duarte Mendes*[1],[2], *Elaine Marcílio Santos*[1],[2], *Marcela Leticia Leal Gonçalves*[1],[2]

(1) Postgraduation Program in Health and Environment, Universidade Metropolitana de Santos, Santos, SP, Brazil.
(2) Dentistry College, Universidade Metropolitana de Santos, Santos, SP, Brazil.
(3) Postgraduation Program in Veterinary Medicine in The Coastal Environment, Universidade Metropolitana de Santos, Santos, SP, Brazil.
(4) Postgraduation Program in Biophotonics Applied to Health Sciences, Universidade Nove de Julho, São Paulo, SP, Brazil.

CORRESPONDING AUTHOR: Marcela Leticia Leal Gonçalves, Francisco Glicério Avenue, 8 - Encruzilhada, Santos - SP, ZIP code: 11045-002. Phone: +55 13 32283400. E-mail: marcelalleal@hotmail.com

SUMMARY

Background: In Brazil, traffic accidents have been on the rise. There is a very high social and economic impact in the country, either by the direct sequelae that are left by the trauma, or by the deaths caused by it.
Objective: This study aims to analyze the medical care associated with traffic accident-induced trauma in Baixada Santista, a Brazilian region comprising nine cities (six of which were included in this paper), along with the causes and consequences of such incidents for the population.
Methods: For data collection, DATA SUS BRASIL, which is a governmental website, was used. The impact that trauma causes on people and the main bodily injuries produced in them, the most affected groups and the cities of the Baixada Santista where the accidents occurred were collected. The data were analyzed using descriptive statistics / relative frequencies.
Results: The analysis reveals that young men, particularly those who ride motorcycles, are most susceptible to traffic accidents. In relation to the cities of the Baixada Santista, we verified that the city of Santos emerges as the primary location for these accidents, primarily due to its substantial motorcycle fleet.
Conclusions: There is a pressing need to implement preventive measures targeting young male motorcyclists in this region. It is through awareness that it will be possible to act in the prevention of accidents and reinforce to the public power, not only the need for consciousness, but also the importance of qualifying the nurses and professionals involved in the treatment and rescue of victims.

Keywords: trauma, traffic injuries, prevention, traffic accidents.

INTRODUCTION

The World Health Organization (WHO) regards traffic accidents as predictable, and thus, they are no longer seen as a fatality, but rather as a disease. Traffic accidents are subject to interventions that involve multidisciplinary efforts aimed at their prevention, which means that the leading cause of traumatic death in the world can be significantly reduced or avoided [1]. Traffic injuries result in significant economic losses for individuals, their families, and nations as a whole. These losses arise from both the cost of treatment and the reduced productivity for those disabled due to the sequelae left by the trauma. Additionally, other family members may need to take time off work or school to care for the injured patients. In most countries, traffic injuries cost about 3% of their Gross Domestic Product (GDP), and this figure can go up to 5% in developing countries [2,3]. In 2021, in Geneva, the WHO launched the Decade of Action for Road Safety 2021-2030, with the ambitious goal of preventing at least 50% of deaths and injuries in road traffic by 2030.

DOI: 10.54103/2282-0930/23370
Accepted: 1st July 2024

Every year, the lives of approximately 1.2 million people worldwide are cut short as a result of traffic accidents. Additionally, between 20 and 50 million more people suffer non-fatal injuries [1].

There are several risk factors for Road Traffic Injuries (RTIs), such as the average speed increases, driving under the influence of alcohol or any psychoactive substance or drug and distracted driving caused, for example, by cell phones. After the occurrences, care for injuries is extremely time-sensitive: delays of minutes can make the difference between life and death. Enhancing post-RTI care necessitates timely access to pre-hospital services and the improvement of both pre-hospital and hospital care quality through specialized training programs [4].

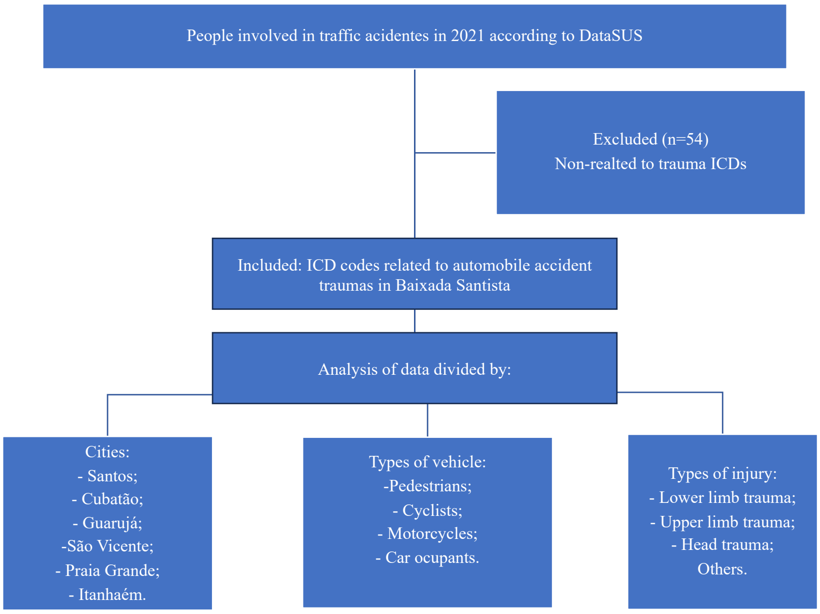
Road traffic injuries are a global health challenge. The number of road traffic deaths continues to rise steadily, from 1.15 million in 2000 to 1.35 million in 2018. Of the 56.9 million deaths worldwide, road traffic injuries account for about 2.37% and are the eighth cause of global death [1,5].

In Brazil, RTIs have been on the rise, following the global trend. In the year 2020, 32,000 deaths were caused. The cumulative numbers from 1980 to 2010 show that almost one million deaths were recorded and this statistic reached one million and three hundred thousand deaths in 2020. As early as 1990, it was already envisioned that, if appropriate measures were not taken, injuries resulting from traffic accidents would become the third leading cause of death by 2020. In 1998, the new Brazilian Traffic Code, governed by Law No. 9,503, came into effect, regarded as the hope for reducing the increasing number of ITRs. However, the new laws, municipal traffic control, vehicle safety improvements, and electronic enforcement have not succeeded in significantly reducing deaths or disabilities resulting from trauma [6].

Baixada Santista is comprised of 9 cities. These are Santos, São Vicente, Cubatão, Guarujá, Praia Grande, Itanhaém, Peruíbe, Mongaguá, and Bertioga. Our study will focus on the first six cities in the Baixada, as the latter three did not have sufficient data related to traffic accidents and were therefore excluded from the study. Santos is the most populous city, with 418,608 inhabitants, and has the largest vehicle fleet, consisting of 139,336 motorcycles and 80,231 automobiles. Itanhaém is the city with the largest territorial area, but the smallest vehicle fleet, with 30,970 automobiles and 14,147 motorcycles. Cubatão is the city with the smallest territorial area, covering 142,879 square kilometers. It has one of the smallest vehicle fleets, 14,547 motorcycles and 30,786 automobiles. The city of Guarujá has the third-largest vehicle fleet, with 67,608 automobiles and 65,135 motorcycles, and it also has the third-highest population density in Baixada Santista, with 1,986.73 inhabitants per square kilometer. The city of Praia Grande has the second-largest population in the Baixada region, with 349,935 inhabitants. São Vicente has the fourth largest vehicle fleet, consisting of 91,064 automobiles and 14,013 motorcycles [7].

This article aims to analyze medical care related to traffic accidents, according to hospitalizations in the hospital network (public or private), in the year 2021, in Baixada Santista. With this data, a better understanding of traffic accidents that occur in this region could be achieved, so that prevention measures could be taken.

Figure 1. Study flowchart.



METHODS

Study design

This study is characterized as: retrospective, cross-sectional and observational. The study flowchart is shown in Figure 1.

Studied population

The population studied in this research comprises patients of all ages, both men and women, from Baixada Santista, involved in traffic accidents and treated in both public and private hospitals.

The year 2021 was chosen as it was the year following the onset of the COVID-19 pandemic, which affected the entire global population. Therefore, we excluded the year 2020. Due to people staying in their homes and fewer vehicles on the roads and highways, the study would be compromised in its evaluation. In 2021, the vehicle movement returned to normalcy.

Inclusion criteria

Patients involved in traffic accidents, whose ICD codes (International Classification of Diseases) were related to automobile accident traumas, occurring in Baixada Santista, in the year 2021.

Exclusion criteria

Fifty-four patients were excluded from this study (representing a total of 4.7% of the sample of 1,137 patients). These exclusions account for 0.88% of women and 3.8% of men in the total studied sample, who were hospitalized in hospitals, whose primary hospitalization ICD code (International Classification of Diseases) was not directly related to trauma caused by a traffic accident. They had the following codes: T81.3 (35 patients), R02 (3 patients), T88.8 (2 patients), T85.8 (1 patient), T84.6 (1 patient), T84.0 (1 patient), T81.4 (1 patient), L97 (2 patients), J90 (2 patients), Z47.0 (2 patients), C43.0 (1 patient), M84.1 (1 patient), M20.2 (1 patient), and M19.1 (1 patient).

Data collection

The procedure used in this research was data collection through the DATASUS BRASIL website, ftp://ftp.datasus.gov.br/dissemin/publicos/SIHSUS/200801_/ Dados/ [8], accessing the DATASUS FTP server and downloading the files with the specified extension.dbc (RDUFYYMM.DBC) containing SIH data for each city, in the year of 2021.

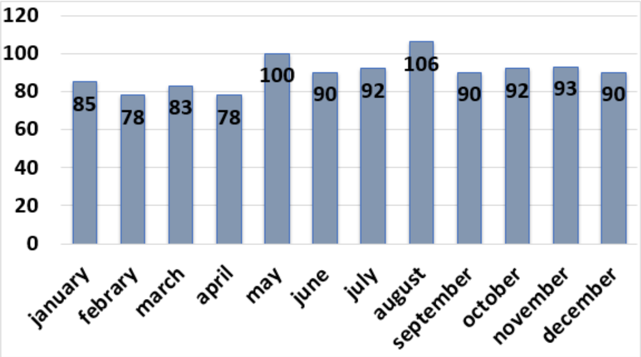
Data analysis

Hospitalizations both by ICD and by the location of the accident occurrence in different cities in Baixada Santista were categorized. The mechanisms that led to the accidents, stratifying them into traffic accidents involving pedestrians, bicycles, motorcycles, or automobiles were also analyzed. Therefore, a comprehensive view of the accident mechanisms and their physical consequences for the studied population could be achieved. The diseases caused by the vast array of different trauma mechanisms were also assessed. The data was presented through descriptive statistics, in relative frequencies.

RESULTS

Firstly, the distribution of hospital admissions due to traffic trauma in Baixada Santista in general, in the year 2021, divided into the number of admissions per month, was analyzed. These results are presented in Figure 2.

Figure 2. Distribution of hospital admissions in Baixada Santista in the year 2021, divided into the number of admissions per month.



Pedestrians

Pedestrians comprised the smallest group, with 103 patients (9.5%) out of all analyzed patients (1,077), consisting of 76% men and 24% women. Despite this, due to their physical vulnerabilities during a traffic accident, they proportionally suffered more from lower limb traumas (the primary point of contact between the pedestrian and the colliding vehicle, regardless of the vehicle type). There was a total of fifty-four lower limb fractures, which accounts for 52.40% of the group of 103 pedestrians traumatized throughout the year 2021. In second place were Traumatic Brain Injuries (TBI), with twenty-one patients hospitalized due to collisions with vehicles, representing 20.4% of all analyzed pedestrians, followed by upper limb injuries in only twelve patients (11.7%). It is easy to understand that during a pedestrian fall, the likelihood of their head striking the ground or object colliding with them is very high. This leads to severe injuries, as the head lacks any protection. The age group of the pedestrian group, as well as all the other groups studied, was divided into three groups: Group 1 included ages 0 to 15, Group 2 included ages 16 to 54, and Group 3 included individuals over 55 years old. Group 1 represented 9.7% of the patients, Group 2 represented 53.4%, and Group 3 represented 36.9% of the analyzed patients. Regarding the location of accidents involving pedestrians, only 9 (8.7%) accidents occurred in Cubatão, 19 (18.5%) in Santos, 3 in Praia Grande (2.9%), 55 (53.4%) in São Vicente, 13 (12.6%) in Itanhaém, and 4 (3.9%) in Guarujá.

Cyclists

Cyclists accounted for 15.3% of all analyzed patients, totaling 165 accident victims. Among them, 24 (14.5%) suffered from Traumatic Brain Injury (TBI). The most common injury among cyclists was lower limb trauma, with a total of 52 fractures in patients, representing 31.5% of cases. Upper limb traumas were present in 32 (19.4%) patients. Clavicle fractures were present in 11 patients (6.6%). Analyzing these data, it becomes clear that lower limb fractures occur more frequently, as this is the most affected area of the body in falls and collisions with automobiles. On the other hand, traumatic brain injury was statistically less common than in pedestrians, as cyclists often use helmets, which protect them against head injuries, while pedestrians do not have any head protection.

The most common geographic location of accidents involving cyclists was in the city of Itanhaém, with 79 traumatized cyclists requiring hospitalization (47.9%), followed by the city of Santos with 51 patients (30.9%), and then Cubatão with 13 (7.9%), Guarujá with 12 (7.3%), São Vicente with 10 (6.0%), and Praia Grande recorded no accidents involving cyclists. This group consisted of 121 males (73.3%) and forty-four females (26.7%). Regarding the patients' age groups, Group 1 represented patients aged 0 to 15 years, with twenty-four patients (14.5%), Group 2 represented patients aged 16 to 54 years, with 116 patients (70.3%), and Group 3 represented patients over 55 years old, with 25 patients (15.2%). In the group of cyclists, there were no hospitalized patients who suffered fatalities.

Motorcyclists

Motorcyclists accounted for the largest analyzed group, totaling 728 patients admitted to hospitals in Baixada Santista due to traffic accidents. Out of this group, fifty-four patients were excluded as their admission ICD was not compatible with trauma. This group represented 67.6% of all analyzed patients. It is evident that motorcyclists are more susceptible to traffic accidents. Men represented 83.2% of this group (606), and women 16.8% (122). The age group was divided into three groups, similar to pedestrians and cyclists. Group 1 (0 to 15 years old) had five patients (0.7%), Group 2 was composed of 678 patients (93.1%), and Group 3 consisted of 45 patients (6.2%). The group of motorcyclists presented a distribution of the main body injuries as follows: 354 patients with lower limb trauma (48.6%), 151 patients with upper limb trauma (20.7%), sixty-four patients with head trauma (8.8%). Regarding the location of the accidents, the geographical distribution was as follows: in Santos, there were 288 accidents (39.6%), in Itanhaém, there were 185 accidents (25.4%), followed by São Vicente with 106 accidents (14.6%), Guarujá with 88 (12.0%), Cubatão with 34 (4.7%), and finally, the municipality of Praia Grande with 27 (3.7%). There were sixteen recorded deaths, corresponding to 2.2% of the patients studied in this sample.

Car and truck occupants

The last group analyzed was car and pickup truck occupants, representing eighty-one patients (7.5%). This group consisted of sixty-three men (77.7%) and eighteen women (22.3%). The age group was distributed as follows: group 1 (0-15 years) with seven patients (8.7%), group 2 (16-54 years) with fifty-eight patients (71.6%), and group 3 (55 years or older) with sixteen patients (19.7%). The most common traumas were lower limb injuries with thirty-eight patients (47.5%), traumatic brain injuries in twelve patients (15%), upper limb fractures in eight cases (10%), abdominal traumas in seven cases (8.8%), and finally, facial traumas in five cases (6.3%). As for the geographical location of the accidents, twenty-seven occurred in São Vicente (33.3%), twenty in Santos (24.7%), twelve in Cubatão (14.8%), eleven in Praia Grande (13.6%), six in Guarujá (7.4%), and five in Itanhaém (6.2%).

The gender and ages distribution of accidents, divided by means of transportation, can be observed in Figure 3. The most common types of bodily injuries (lower limb fractures, upper limb fractures, and traumatic brain injuries) are presented in Figure 4. The geographical distribution of accidents, by city in the Baixada Santista, is presented in Figure 4, 5. Total victims and fatalities are represented in Figure 6.

Figure 3: Distribution by gender (pedestrians - blue, cyclists - green, motorcyclists - gray and car occupants- yellow) and age (0 to 5 years – blue, 16 to 54 years – green; 55 years or more – gray) of accidents, divided by groups.

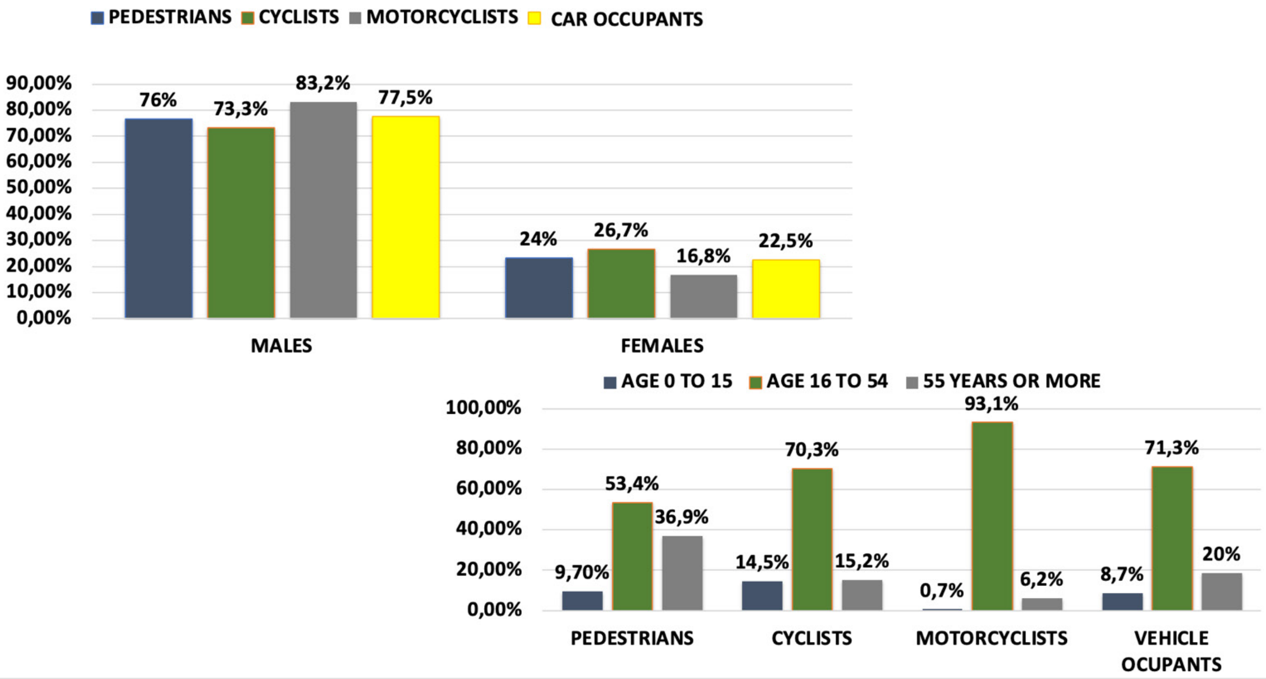


Figure 4: Injuries in traffic accidents in Baixada Santista in 2021, by groups. The blue color represents lower limb injuries, orange represents upper limb injuries, and the gray color represents head and brain injuries, in pedestrians, cyclists, motorcyclists and vehicle occupants.

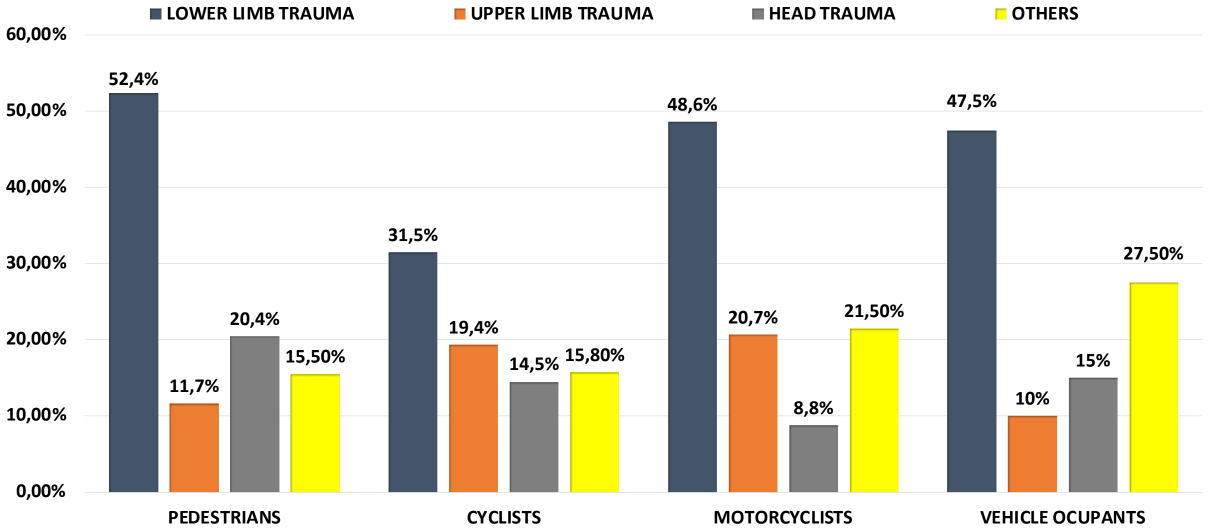


Figure 5: Geographical distribution of accidents, by city (Santos, Cubatão, Guarujá, Itanhaém, Praia Grande and São Vicente) in Baixada Santista. The blue color represents pedestrians, orange represents cyclits, the gray color represents motorcyclists, and yellow represents vehicle occupants.

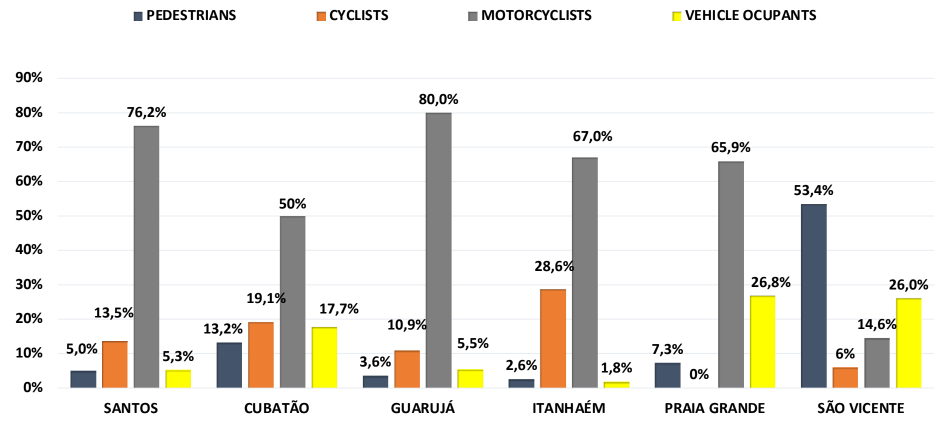
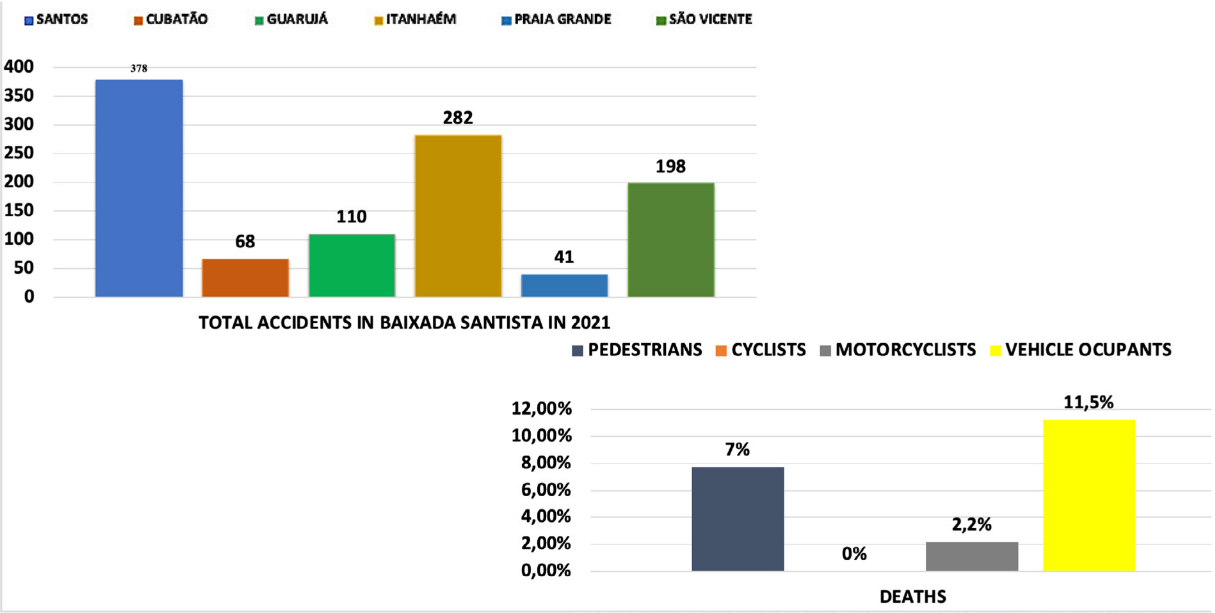


Figure 6: Incidence of traffic accident victims in Santos (light blue), Cubatão (orange), Guarujá (green), Itanhaém (dark yellow), Praia Grande (blue) and São Vicente (green), and deaths during hospitalizations of traffic accident victims that occurred in Baixada Santista in 2021 (the blue color represents pedestrians, orange represents cyclits, the gray color represents motorcyclists, and yellow represents vehicle ocupants).



DISCUSSION

Traffic accidents follow a trimodal curve regarding patient mortality. The first peak of mortality is related to the severity of the accident at the scene. The second peak is linked to pre-hospital care and its quality; to the time involved between the care and the patient's transfer to the hospital unit where they will be admitted. The third peak is associated with complications arising from trauma-induced injuries.

We can only intervene in the first peak with awareness campaigns involving the population on how to prevent accidents, emphasizing the importance of seatbelt use for all vehicle occupants, ensuring children are properly secured in age-appropriate safety seats, promoting respect for traffic signs and speed limits on public roads, and advocating for the use of protective gear for motorcyclists (closed helmets with visors, jackets, protective pants, and closed-toe footwear).

During the second peak, intervention can be achieved by providing high-quality, specialized care delivered by qualified trauma care specialists. This requires ambulances with advanced extrication tools and life support equipment for pre-hospital care.

In the third peak, high-quality hospital care can be provided with all the necessary support equipment for performing complex surgeries and modern diagnostic equipment, such as CT scans available in all hospitals, and the use of FAST (Focused Assessment with Sonography for Trauma) in trauma rooms. Acute treatment for polytrauma patients is administered by specialized trauma care professionals [9].

The way accidents occur, depending on the environment, type of vehicle, whether the driver was wearing a seatbelt or was speeding, and the mechanism of the accident (such as pedestrian collision, motorcycle or car collision, or involvement of bicycles), is of great importance. Each mode has its peculiarities regarding the type and severity of organic injury inflicted. This directly impacts both patient mortality and survival, as well as the recovery time and resulting sequelae. This is because each type of injury depends on the intensity of the trauma and the exposure of our body to that trauma (Figure 4) [10-12].

Analyzing the factors leading to a traffic accident, William Haddon Jr, an American epidemiologist, was a pioneer in classifying traffic accidents as an epidemic. In his 1980 article, he examined the factors related to accidents and developed a matrix where he divided these factors into phases. The phases include pre-event (what occurs before the event and facilitates its occurrence), the event itself (the accident), and post-event (after the accident has occurred). Haddon further subdivides the factors that interfere with each phase, such as human factors, the causative agent, and environmental factors (physical and socio-cultural). This is of great significance because it allows us to understand the mechanisms involved before, during, and after the accident, enabling us to take action in each phase to minimize the impacts of traffic incidents. Furthermore, establishing campaigns focused on prevention and reduction of accidents, directly targeting all three phases, leads to a reduction in accidents (the primary objective), as well as an improvement in the care provided to accident victims and a reduction in the resulting sequelae [12,13].

The factors related to traffic incidents can be divided into: predisposing factors (such as gender, age, marital status, level of education), facilitating factors (such as income level, healthcare access conditions), precipitating factors (such as road and highway maintenance, driver distraction, and vehicle fleet maintenance quality), and reinforcing factors like lack of awareness while driving, dangerous driving, obesity, and advanced age (9-14). Patients without protection, such as, for example, not wearing helmets for motorcyclists and cyclists or not using seat belts for car occupants, are more likely to sustain severe injuries at the time of the accident [10,14].

The type of injury is directly related to the mode of transportation and its speed at the time of the accident (Figure 3); because the higher the speed and energy of the impact, the greater the trauma inflicted on the body of individuals, as the energy formula is $E = m \times c^2$ (energy equals mass (m) times the velocity (c) squared). Therefore, speed is the variable that exerts the most influence on the energy of the trauma, and hence, we must control it through active and passive enforcement [15].

Observing the dynamics of traffic incidents worldwide, we realize that the most exposed group to trauma remains the same, consisting of men between the ages of 18 and 60 who are motorcycle users [16]. We believe the main reasons for the prevalence of this population lies in the fact that many men work with their motorcycles, making deliveries, which usually have a time deadline. So many of them end up speeding or crossing red lights in order to make it to destinations on time. Moreover, young men usually seek the adrenaline of high speed, even if not for a particular reason. The installation of more speed cameras and the full use of personal protective equipment could help to avoid accidents and, if they do occur, the risk would be lower.

Socio-economic factors, such as income levels, that sometimes make it necessary for the population to work with their cars / motorcycles, urbanization, and healthcare accessibility, might influence the incidence and severity of traffic-related injuries in Baixada Santista. Healthcare accessibility outside of the public health system is expensive, and therefore, not accessible to a great part of the population. Consequently, a lot of public health investments need to go the treatment of accident sequalae, and some of these accidents could be prevented thorough caution and awareness.

After a thorough study of the accidents that occurred in Baixada Santista, and the stratification of the causes leading to traffic accidents and understanding how we can address this issue, we observe how it is possible to act in the prevention of accidents by controlling speed on the roads and improving the quality of care, both at the scene of the accident and in hospital care [10,18]. The government plays a crucial role in this matter, as the development of accident prevention policies and the enforcement of traffic laws, including speed control on roads, are the cornerstone of accident reduction [17,18]. In the city of Santos, the Traffic Engineering Company (CET) is aligned with the Sustainable Development Goals, aiming to enhance this aspect in the city [19]. It is crucial for public health policies like this to be aligned in order to enable better care, both pre-hospital and intra-hospital.

Although this study is limited to the results of datasets of a specific region, its finding can be applied to general population, seeing as the results seem to show similar characteristics worldwide. A study in Nigeria showed that the leading causes of road traffic accidents there are human factors; speed violation, loss of vehicle control and dangerous driving. They agree with our study, emphasizing that the accidents are preventable, and that sensitization and enforcement of safe road principles among commercial vehicles and car drivers would help curb this menace. The authors also believe that government at all levels should implement strong policies aimed at reducing the speed of vehicles on roads [20]. A survey that was conducted in China examined 234 major road traffic accidents recorded in 27 Chinese provinces from 1997 to 2014. They analyzed the relationships among the contributing factors. At the preconditions for unsafe acts level, "visual limitation", "fatigue driving," and "vehicle faults" were strong predictors [21].

After analyzing all the data collected in this study, we conclude that the dynamics of traffic accidents occurring in Baixada Santista follow common causes and mechanisms seen worldwide. Young male motorcyclists constitute the most affected demographic group. Imprudent driving, particularly distraction and excessive speed, are the main causes of accidents and fatalities. It is crucial to focus on prevention through awareness campaigns, but enforcement and punishment for speeding violations need to be carried out. A video campaign was released after this research was conducted, and it available on YouTube [22], to show data and make people more aware of preventive measures. Specific infrastructure improvements, such as better road lighting, dedicated motorcycle lanes, and improved pedestrian crossings should also be made. Policy changes, such as stricter enforcement of traffic laws, mandatory helmet use, and targeted public awareness campaigns should be more and more conducted. Finally, improving medical care, both pre-hospital and hospital, is essential to minimize the sequelae produced by traffic incidents.

Data availability statement

The datasets used in this study are from public domain, available from DATASUS, a Brazilian government website. Available from https://datasus.saude.gov.br/.

REFERENCES

1. World Health Organization –Global status report on road safety: time for action. Geneva, World Health Organization, Available from: www.who.int/violence_ injury_prevention/road_safety_status/.
2. Andrade, SSCA, Jorge, MHPM. Hospital admissions for injuries resulting from land transport accidents in Brazil, 2013: length of stay and expenses. Epidemiology and Health Services [online]. 2017. Accessed December 4, 2022, pp. 31-38.
3. Salvarani CP, Colli BO, Carlotti Júnior CG. Impact of a program for the prevention of traffic accidents in a Southern Brazilian city: a model for implementation in a developing country. Surg Neurol. 2009 Jul;72(1):6-13; discussion 13-4.
4. Cociu S, Ioncz O, Ciobanu D, Cebanu S. Knowledge, and attitudes regarding road safety among drivers. A health risk management. March 11, 2023;4(2):25-32.
5. Chang FR, Huang HL, Schwebel DC, Chan AHS, Hu GQ. Global road traffic injury statistics: Challenges, mechanisms and solutions. Chin J Traumatol. 2020 Aug;23(4):216-218.
6. Bacchieri G, Barros AJ. Traffic accidents in Brazil from 1998 to 2010: many changes and few effects. Public Health Journal. 2011 Oct;45(5):949-63. English, Portuguese. Epub 2011 Sep 16.
7. IBGE – BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS. Cities Portal. IBGE, 2023. Available at: https://ibge.gov.br/cidades-e-estados/sp.
8. Platform DATASUS. Available from https://datasus.saude.gov.br/ (2020).
9. Qiu J, Yang A, Li K, Zhao H, Qin M. Analysis on alteration of road traffic casualties in western China from multi-department data in recent decade. Front Public Health. 2022 Nov 10;10:972948.
10. Razzaghi A, Soori H, Kavousi A, Abadi A, Khosravi A, Alipour A. Risk factors of deaths related to road traffic crashes in World Health Organization regions: A systematic review. Arch Trauma Res 2019; 8:57-86. Received: 13-07-2019, Accepted: 21-08-2019, Web Publication: 07-10-2019.
11. Se C, Champahom. T, Wisutwattanasak P, Jomnonkwao S, Ratanavaraha V. Temporal instability and differences in the severity of injuries between restrained and unrestrained drivers in speed-related accidents. Sci Rep. 2023 Jun 16;13(1):9756.
12. Bocage C, Mashalla Y, Motshome P, Fane O, Masilo-Nkhoma L, Mathiba O, Mautle E, Kuiperij B, Mmusi T, Holmes JH, Tam V, Barg FK, Wiebe DJ. Applying the Haddon matrix conceptual model to guide motor vehicle crash injury research and prevention in Botswana. Afr J Emerg Med. 2020;10(Suppl 1):S38-S43.
13. Haddon W Jr. Advances in the epidemiology of injuries as a basis for public policies. Public Health Rep. 1980 Sep-Oct.;95(5):411-21.
14. Kambiz M, Arash F, Hassan B, et all. Effective Factors in Severity of Traffic Accident-related Traumas; an Epidemiologic Study Based on the Haddon Matrix. Emerg (Tehran) 2016 Spring; 4(2): 78-82.
15. Gasana J, Albahar S, MAlkhalidi M, Al-Merkhled Q, El Reda et all. Risky Roads in Kuwait: An Uneven Toll on Migrant Workers. Int J Environ Res Public Health. 2022 Aug 7;19(15):9726.
16. Ospina-Mateus H, Garcia S B, et all. Dataset of traffic accidents in motorcyclists in Bogotá, Colômbia. Data Brief. 2022 Jul 16;43:108461.
17. Sadeghi-Bazargani H, Saadati MSpeed Management Strategies: A Systematic Review. Bull Emerg Trauma. 2016 Julho;4(3):126-33.
18. Cascetta E, Punzo V, Montanino M. Empirical Analysis of the Effects of Automated Section Speed Enforcement System on Traffic Flow at Highway Bottlenecks. Transportation Research Record: Journal of the Transportation Research Board. 2011; 2260:83–93.
19. Santos 2030 ODSs Portal. Available at: https://www.santos.sp.gov.br/?q=portal/ods-santos-2030.
20. Anebonam U, Okoli C, Ossai P, Ilesanmi O, Nguku P, Nsubuga P, Abubakar A, Oyemakinde A. Trends in road traffic accidents in Anambra State, South Eastern Nigeria: need for targeted sensitization on safe roads. Pan Afr Med J. 2019 Jan 25;32(Suppl 1):12. doi:
21. Zhang Y, Jing L, Sun C, Fang J, Feng Y. Human factors related to major road traffic accidents in China. Traffic Inj Prev. 2019;20(8):796-800.
22. Available from: https://www.youtube.com/watch?v=xGpXi2ddsWo&t=4s.

Effect of COVID-19 on Cardiovascular Disease Mortality in a Medium-Sized City

Gilberto Campos Guimarães Filho[1]

[1] Federal University of Jataí

CORRESPONDING AUTHOR: Gilberto Campos Guimarães Filho, Federal University of Jataí. Email: camposguimaraes@yahoo.com.br

SUMMARY

Introduction: Patients hospitalized for COVID-19 have a high prevalence of cardiovascular risk factors, such as hypertension and diabetes mellitus, in addition to chronic cardiovascular conditions, such as ischemic heart disease and heart failure. Other fatal events have also occurred, and are somehow indirectly associated with the pandemic, such as deaths from neglected or inadequately treated diseases due to an overburdened health system or fear of leaving home at the height of the COVID-19 pandemic. The present study aims to evaluate the impact of COVID-19 on CVD statistics in a medium-sized city.
Methods: A retrospective observational study with 546 patients who died due to cardiovascular diseases, between January 1, 2019 and December 31, 2020. Regarding the year of death, continuous variables were compared using the unpaired t-test and categorical variables using Pearson's chi-square or Fisher's exact test. Considering a significance level of 0.05 in the two-tailed test.
Results: A total of 545 deaths due to cardiovascular diseases were evaluated, 272 in 2019 and 274 cases in 2020. There was no difference in age and sex ratios between the years evaluated. There was a higher frequency of deaths at home in 2020 compared to 2019; a reduction in the frequency of deaths occurring in hospitals.
Conclusion: The results indicate an increase in the number of deaths at home due to CVD reported by SIM in 2020 compared to the same period in the previous year, in a medium-sized Brazilian city.

Keywords: Covid 19; Cardiovascular Diseases; Death; Myocardial infarction; Cardiac events.

INTRODUCTION

On March 11, 2020, the World Health Organization (WHO) declared the public health situation involving the disease named COVID-19, caused by the novel coronavirus SARS-CoV-2, a pandemic [1]. It causes a severe acute respiratory syndrome, which invades cells via the angiotensin-converting enzyme 2 [2] and has been responsible for approximately 713,000 deaths in Brazil [3].

Patients hospitalized due to COVID-19 exhibit a high prevalence of cardiovascular risk factors, such as hypertension and diabetes mellitus, as well as chronic cardiovascular conditions, such as ischemic heart disease and heart failure [4-8], contributing to a significant increase in COVID-19-related mortality [9]. Additionally, other fatal events have occurred, indirectly associated with the pandemic, such as deaths from acute myocardial infarction and stroke due to neglected or inadequately treated hypertension, dyslipidemia, and diabetes because of the overwhelmed healthcare system or fear of leaving home at the height of the COVID-19 pandemic [9,10].

Considering that cardiovascular diseases (CVD) are the leading cause of death in Brazil [11], this study aims to evaluate the impact of COVID-19 on CVD mortality statistics in a medium-sized city.

DOI: 10.54103/2282-0930/25501
Accepted: 20th September 2024

METHODS

Study Type and Location

This is an observational, retrospective study conducted in the city of Rio Verde, Goiás, Brazil, a medium-sized city with an estimated population of nearly 248,000 [12].

Population, Sample, and Sampling

The study comprised 546 patients who died from cardiovascular diseases between January 1, 2019 (pre-COVID-19 pandemic) and December 31, 2020 (during the COVID-19 pandemic). Deaths where the primary cause was not cardiovascular disease were excluded.

Data

The analysis was based on secondary ordinary administrative data from the SUS Mortality Information System (SIM). The International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) codes were used to identify the underlying causes of death due to ischemic heart disease (I20-I25), heart failure (I50), hypertensive diseases (I10-I15), arrhythmias (I49), and other conditions. To examine the potential indirect effect of the pandemic, deaths with COVID-19 as the underlying cause were excluded from the analysis.

Statistical Analysis

Categorical variables were summarized using absolute (n) and relative (%) frequencies, while continuous variables were summarized using means and standard deviations (SD). An outlier in age (13 years) was removed due to the non-cardiac cause of death but did not compromise the statistical analysis.

Regarding the year of death, continuous variables were compared using the unpaired t-test, and categorical variables were compared using Pearson's chi-square or Fisher's exact test. All analyses were performed using R software version 4.3.2 (R Core Team, Vienna, Austria), with a significance level of 0.05 for two-tailed tests.

Ethical Considerations

Approval from a Human Research Ethics Committee was not required due to the use of non-identified, publicly available data.

RESULTS

A total of 545 deaths due to cardiovascular diseases were evaluated, with 272 cases in 2019 (pre-pandemic) and 274 cases in 2020 (during the pandemic). There were no differences in age and sex proportions between the years evaluated (Table 1).

An increased frequency of deaths at home was observed in 2020 compared to 2019. In this regard, there was a reduction in the frequency of deaths occurring in hospitals or other healthcare institutions (p<0.001). For other diseases, groups, and subgroups of diseases, no differences in proportions were observed.

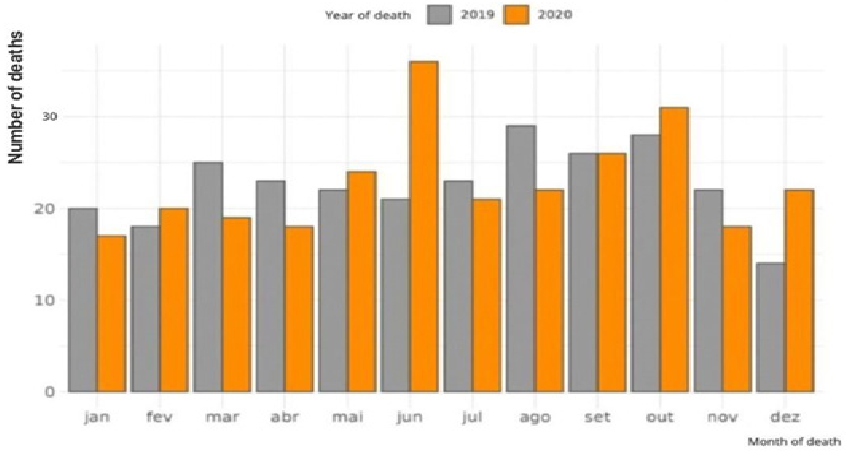
Table 1. Epidemiological Profile of Cardiovascular Disease Deaths in Rio Verde/GO: January 2019 to December 2020

Variables	**2019**, N = 271	**2020**, N = 274	**p**
Age (years)	69.88 (15.24)	70.88 (14.98)	**0.54**
Age Range			**0.48**
21-59 years	64 (23.62%)	54 (19.71%)	
60-75 years	94 (34.69%)	105 (38.32%)	
76-102 years	113 (41.70%)	115 (41.97%)	
Sex			**0.36**
Female	113 (41.70%)	125 (45.62%)	
Male	158 (58.30%)	149 (54.38%)	
Place of Death			**<0.001**
Home	48 (17.71%)	89 (32.48%)	
Hospital	145 (53.51%)	133 (48.54%)	
Others	6 (2.21%)	5 (1.82%)	
Other Health Institutions	71 (26.20%)	45 (16.42%)	
Public Place	1 (0.37%)	2 (0.73%)	
MI			**0.52**

No	189 (69.74%)	184 (67.15%)	
Yes	82 (30.26%)	90 (32.85%)	
DM			**0.13**
No	237 (87.45%)	227 (82.85%)	
Yes	34 (12.55%)	47 (17.15%)	
HT			**0.75**
No	204 (75.28%)	203 (74.09%)	
Yes	67 (24.72%)	71 (25.91%)	
Dyslipidemia			**>0.99**
No	267 (98.52%)	270 (98.54%)	
Yes	4 (1.48%)	4 (1.46%)	
Obesity			**0.070**
No	267 (98.52%)	263 (95.99%)	
Yes	4 (1.48%)	11 (4.01%)	
Heart Disease			**>0.99**
No	266 (98.15%)	269 (98.18%)	
Yes	5 (1.85%)	5 (1.82%)	
HF			0.37
No	270 (99.63%)	270 (98.54%)	
Yes	1 (0.37%)	4 (1.46%)	
Arrhythmia			**>0.99**
No	267 (98.52%)	270 (98.54%)	
Yes	4 (1.48%)	4 (1.46%)	
Mental Disorder (Abusive Smoking)			**0.45**
No	253 (93.36%)	260 (94.89%)	
Yes	18 (6.64%)	14 (5.11%)	
Mental Disorder (Abusive Alcohol)			**0.75**
No	266 (98.15%)	270 (98.54%)	
Yes	5 (1.85%)	4 (1.46%)	
Nephropathy			**>0.99**
No	266 (98.15%)	269 (98.18%)	
Yes	5 (1.85%)	5 (1.82%)	
Respiratory Tract Diseases			**0.37**
No	258 (95.20%)	265 (96.72%)	
Yes	13 (4.80%)	9 (3.28%)	
Endocrine-Metabolic Diseases			**0.22**
No	264 (97.42%)	271 (98.91%)	
Yes	7 (2.58%)	3 (1.09%)	
Depressive Episode			**>0.99**
No	271 (100.00%)	273 (99.64%)	
Yes	0 (0.00%)	1 (0.36%)	
Senility			**0.76**
No	264 (97.42%)	268 (97.81%)	
Yes	7 (2.58%)	6 (2.19%)	

MI: myocardial infarction; DM: diabetes mellitus; HT: hypertension; HF: heart failure. Pearson's chi-square test.

Figure 1. Monthly Evolution of Cardiovascular Disease Deaths in Rio Verde (GO) by Year



In both years assessed, individuals who died from cardiovascular diseases had a statistically similar average age compared to those who did not die from this cause (Figure 2).

The population aged 60-75 years had the highest proportion of CVD deaths in both 2019 and 2020 (Figure 3). Additionally, the proportion of CVD deaths was statistically similar between the years, regardless of age group.

Figure 2. Comparison of Average Age of Cardiovascular Disease Deaths in Rio Verde (GO) by Underlying Cause and Year

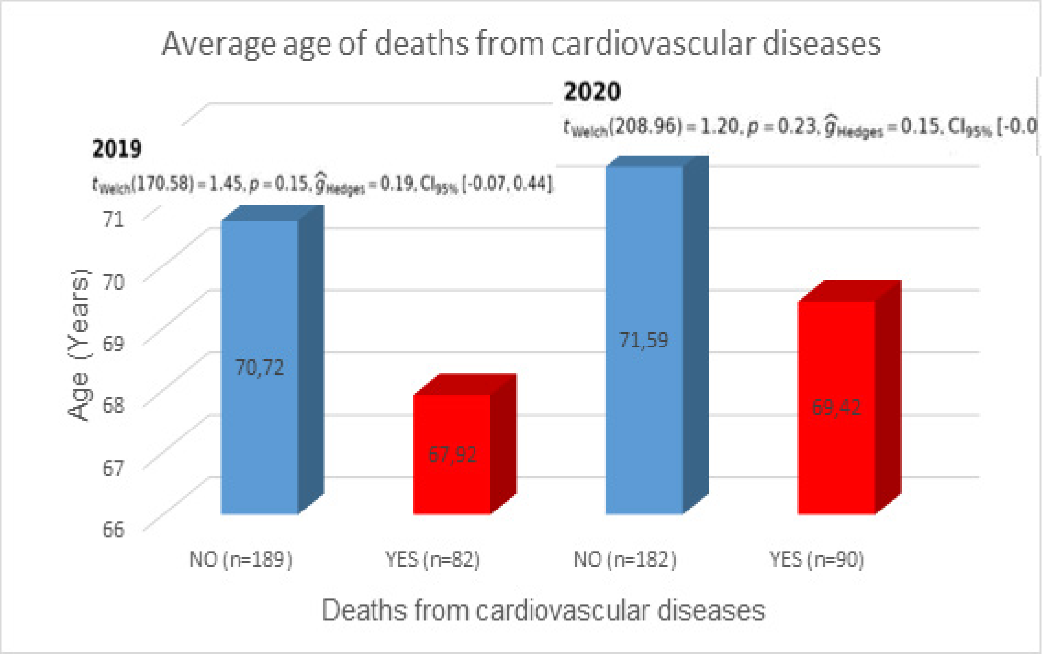
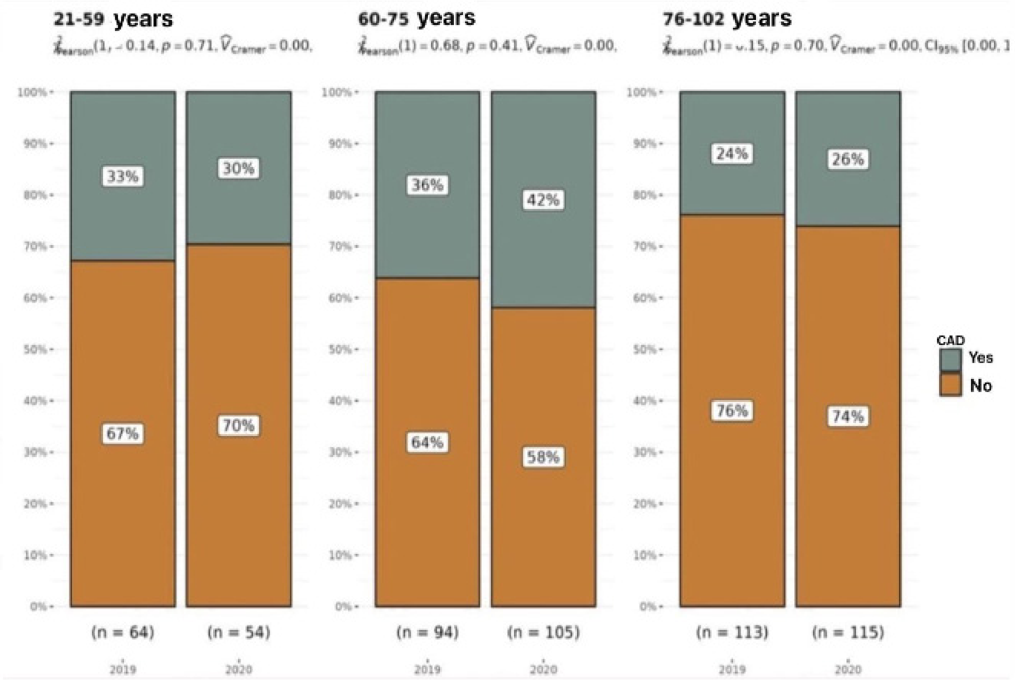
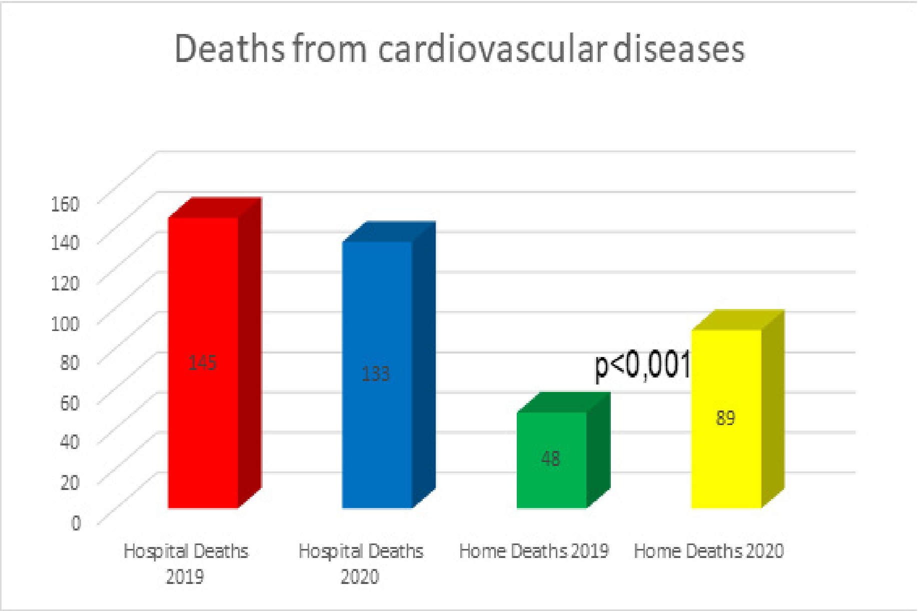


Figure 3. Proportion of Cardiovascular Disease Deaths in Rio Verde (GO) by Age Group and Year



Central Figure. Cardiovascular Disease Deaths in Rio Verde (GO) by Location and Year



DISCUSSION

In this analysis, it was observed that the two evaluated groups (2019 vs. 2020) showed no significant differences concerning age, sex, and other cardiovascular comorbidities such as hypertension, diabetes mellitus, obesity, dyslipidemia, and depression. However, the primary result of this study was an 83% increase in home deaths due to cardiovascular diseases (CVD) in 2020 compared to 2019, with these deaths being more frequent in May, June, October, and December, and myocardial infarction (MI) was the leading cause in almost one-third of the cases. There was no difference in the age range of deaths between the evaluated years, although a higher proportion of CVD deaths was found in the 60-75 year age range. Our results support and extend previous findings [13,14].

Regarding the significant increase in home deaths in 2020, during the period of social isolation due to the pandemic, it is important to note that there was a substantial decline in hospitalizations for acute cardiovascular diseases, such as acute coronary syndrome, stroke, atrial fibrillation, and acute or decompensated heart failure, as well as a considerable proportion of excess deaths not attributed to COVID-19 during the ongoing pandemic [15-25]. Our study showed approximately a 16% reduction in hospitalizations in 2020, with a more pronounced reduction of 25% in May and June, which could be a risk factor for the mismanagement of chronic diseases such as hypertension, dyslipidemia, and diabetes, contributing to adverse cardiovascular outcomes.

Although the COVID-19 pandemic has been an unprecedented threat to global health in recent times [26], reports of potential side effects from widespread strategies to prevent the rapid spread of SARS-CoV-2 and relieve the burden on health systems, with reduced attention to non-communicable chronic diseases, as well as the continuous increase in deaths reported by the media, may have led to fear among this population in seeking adequate maintenance of their treatments, resulting in increased home deaths not due to COVID-19 among patients with established cardiovascular comorbidities. Generally, the reduction in hospital visits for patients with non-communicable chronic diseases, especially hypertension, diabetes, dyslipidemia, and myocardial infarction, was expected as the pandemic necessitated a reorganization of the existing human and technological hospital resources focused on combating COVID-19 [27]. In this study, home deaths due to CVD increased by 83% during the COVID-19 lockdown compared to the previous year. The same result was observed in a cross-sectional study with nearly 3.5 million deaths before and during the pandemic [28]. A divergent result was found in a Danish study with nearly 700,000 patients [29].

The present study showed that the leading cause of death in both evaluated years was myocardial infarction, followed by hypertension and diabetes [11]. Myocardial infarction is the leading cause of death worldwide and requires urgent attention, as does the management of its risk factors such as hypertension and diabetes mellitus. These conditions represent acute events that always require immediate emergency care, even during a public health emergency such as the COVID-19 pandemic. Despite the statistical insignificance of myocardial infarction deaths between the two years evaluated, the significant increase in home deaths due to myocardial infarction in our study may reflect a fear of presenting to the emergency room, having a detrimental impact on patients with myocardial infarction, delaying or avoiding treatment. Similar results were found in other studies that also reported increased mortality from myocardial infarction unrelated to COVID-19 [30,31]. Sometimes, clearer, more frequent, and highly visible communication by healthcare and public health professionals to reinforce the importance of timely medical emergency care and to ensure public safety regarding COVID-19 contamination might have yielded different outcomes.

Another interesting finding from this study was the increase in cardiovascular mortality rates, mainly myocardial infarction, in the months of May, June, October, and December 2020 compared to the same period the previous year. The explanations for this observation are unclear and we can only speculate about this finding, such as the possibility of lower temperatures in the months of May and June increasing the risk of viral infection, which could trigger atherothrombotic processes in patients already suffering from chronic coronary artery disease [32,33]. Additionally, the relaxation of social isolation in October and December, during extended holidays and school vacations, and consequently, an increase in social gatherings and family reunions, might have contributed to the observed increase in deaths. The significant rise in COVID-19 cases in May, June, and October 2020 might also have contributed to the increase in cardiovascular deaths. Given the shortcomings of public cardiovascular prevention policies in the municipality both before and after the pandemic, we might not attribute this variable as a causal factor of the deaths.

Regarding mortality by age group, this study showed a higher proportion of mortality in the 60-75 age group, with no statistically significant difference between the two years evaluated. Although Brazil stands out among the BRICS countries for its successful epidemiological transition, with a rapid reduction in CVD mortality both in cohorts of newborns and in all age groups over time [34], the rapid ageing of the population, the increase in obesity and the decline in healthy diet and physical activity can be major risk factors for cardiovascular death, especially in older and more vulnerable patients [35]. In our study, both groups died of CVD at an average age of 70, which may have contributed to this finding. Excess deaths make it possible to quickly assess the mortality burden directly attributable to COVID-19, as well as its indirect burden, resulting from interruptions in access, use and provision of health services [26,36].

The limitations of our study that deserve to be highlighted are: The first concerns the coverage of the number of deaths in the civil registry, which may have biased our estimates of excess mortality, even though we revised and updated the database. The second refers to the difficulty of analyzing cardiovascular mortality, given the unavailability of this specific information for the entire period analyzed. The third is due to the reliability of the administrative data presented by the SIM, such as incompleteness and the high rate of deaths from ill-defined causes, reflecting on the quality of the completion and processing of death certificates and making it difficult to carry out a more comprehensive analysis considering the population of health insurance beneficiaries who use private health services. Despite the limitations mentioned above, especially in a context such as the pandemic, which imposes significant challenges on the health system, it is worth emphasizing the importance of having a national database which, in global terms, covers around 75% of the Brazilian population and is available relatively quickly. This study adds to others already published in other countries and contributes by examining in detail the effects of the COVID-19 pandemic on cardiovascular mortality in a medium-sized Brazilian city, considering the universe of deaths from COVID-19 and other conditions covered by the SUS.

CONCLUSION

The results indicate an increase in at-home CVD deaths reported by SIM in 2020 compared to the previous year in a mid-sized Brazilian city. Our study underscores the need to reform public health policies to prevent cardiovascular deaths during periods of viral infection spread.

REFERENCES

1. WHO Director-General's opening remarks at the media briefing on COVID-19, 2020a.
2. LAKE, M. A. What we know so far: COVID-19 current clinical knowledge and researchClinical Medicine, Journal of the Royal College of Physicians of London, 2020.
3. https://covid.saude.gov.br/(Painel de casos de doença pelo coronavírus 2019 (COVID-19) no Brasil pelo Ministério da Saúde) [Accessed 02 sept 2024].
4. Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int J Infect Dis. 2020 May;94:91-95. doi: 10.1016/j.ijid.2020.03.017.
5. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ. 2020 May 22;369:m1985. doi: 10.1136/bmj.m1985.
6. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA. 2020 May 26;323(20):2052-2059. doi: 10.1001/jama.2020.6775. Erratum in: JAMA. 2020 May 26;323(20):2098. doi: 10.1001/jama.2020.7681.
7. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020 Aug;584(7821):430-436. doi: 10.1038/s41586-020-2521-4.
8. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020 Apr 7;323(13):1239-1242. doi: 10.1001/jama.2020.2648.
9. Oliveira GMM de, Brant LCC, Polanczyk CA, Malta DC, Biolo A, Nascimento BR, et al.. Estatística Cardiovascular – Brasil 2023. Arq Bras Cardiol [Internet]. 2024;121(2):e20240079. Available from: https://doi.org/10.36660/abc.20240079.
10. Wadhera RK, Shen C, Gondi S, Chen S, Kazi DS, Yeh RW. Cardiovascular Deaths During the COVID-19 Pandemic in the United States. J Am Coll Cardiol. 2021 Jan 19;77(2):159-169. doi: 10.1016/j.jacc.2020.10.055
11. GBD 2021 Demographics Collaborators. Global age-sex-specific mortality, life expectancy, and population estimates in 204 countries and territories and 811 subnational locations, 1950-2021, and the impact of the COVID-19 pandemic: a comprehensive demographic analysis for the Global Burden of Disease Study 2021. Lancet. 2024 May 18;403(10440):1989-2056. doi: 10.1016/S0140-6736(24)00476-8.
12. © 2017 IBGE - Instituto Brasileiro de Geografia e Estatística | v4.6.29.
13. Butt JH, Fosbol EL, Gerds TA, Andersson C, Kragholm K, Biering-Sørensen T, et al. All-cause mortality and location of death in patients with established cardiovascular disease before, during, and after the COVID-19 lockdown: a Danish Nationwide Cohort Study. Eur Heart J. 2021 Apr 14;42(15):1516-1523. doi: 10.1093/eurheartj/ehab028.
14. Janus SE, Makhlouf M, Chahine N, Motairek I, Al-Kindi SG. Examining Disparities and Excess Cardiovascular Mortality Before and During the COVID-19 Pandemic. Mayo Clin Proc. 2022 Dec;97(12):2206-2214. doi: 10.1016/j.mayocp.2022.07.008.

15. Metzler B, Siostrzonek P, Binder RK, Bauer A, Reinstadler SJ. Decline of acute coronary syndrome admissions in Austria since the outbreak of COVID-19: the pandemic response causes cardiac collateral damage. Eur Heart J 2020;41: 1852–1853.
16. Fischer T. Home care in Germany during the COVID-19 pandemic: A neglected population? J Nurs Scholarsh. 2023 Jan;55(1):215-225. doi: 10.1111/jnu.12851.
17. Holt A, Gislason GH, Schou M, Zareini B, Biering-Sørensen T, Phelps M, et al. New-onset atrial fibrillation: incidence, characteristics, and related events following a national COVID-19 lockdown of 5.6 million people. Eur Heart J 2020;41:3072–3079.
18. Vosko I, Zirlik A, Bugger H. Impact of COVID-19 on Cardiovascular Disease. Viruses. 2023 Feb 11;15(2):508. doi: 10.3390/v15020508.
19. Mesnier J, Cottin Y, Coste P, Ferrari E, Schiele F, Lemesle G, et al. Hospital admissions for acute myocardial infarction before and after lockdown according to regional prevalence of COVID-19 and patient profile in France: a registry study. Lancet Public Health 2020;5:e536–e542.
20. Pepera G, Tribali MS, Batalik L, Petrov I, Papathanasiou J. Epidemiology, risk factors and prognosis of cardiovascular disease in the Coronavirus Disease 2019 (COVID-19) pandemic era: a systematic review. Rev Cardiovasc Med. 2022 Jan 17;23(1):28. doi: 10.31083/j.rcm2301028.
21. Boulos PK, Freeman SV, Henry TD, Mahmud E, Messenger JC. Interaction of COVID-19 With Common Cardiovascular Disorders. Circ Res. 2023 May 12;132(10):1259-1271. doi: 10.1161/CIRCRESAHA.122.321952.
22. Solomon MD, McNulty EJ, Rana JS, Leong TK, Lee C, Sung S-H, et al. The Covid-19 pandemic and the incidence of acute myocardial infarction. N Engl J Med 2020;383:691–693.
23. Siregar KN, Kurniawan R, Nuridzin DZ, BaharuddinNur RJ, Retnowati, Handayani Y, Rohjayanti, Halim L. Strengthening causes of death identification through community-based verbal autopsy during the COVID-19 pandemic. BMC Public Health. 2022 Aug 23;22(1):1607. doi: 10.1186/s12889-022-14014-x.
24. Kansagra AP, Goyal MS, Hamilton S, Albers GW. Collateral effect of Covid-19 on stroke evaluation in the United States. N Engl J Med 2020;383:400–401.
25. Andey AS, Daou BJ, Tsai JP, Zaidi SF, Salahuddin H, Gemmete JJ, et al. COVID-19 pandemic—the bystander effect on stroke care in Michigan. Neurosurgery 2020;87:E397–E399.
26. Ronco D, Matteucci M, Ravaux JM, Kowalewski M, Massimi G, Torchio F, et al. Impact of COVID-19 on incidence and outcomes of post-infarction mechanical complications in Europe. Interdiscip Cardiovasc Thorac Surg. 2023 Dec 5;37(6):ivad198. doi: 10.1093/icvts/ivad198.
27. Portela, M.C., de Aguiar Pereira, C.C., Lima, S.M.L., Andrade, CLT., Martins, M. Patterns of hospital utilization in the Unified Health System in six Brazilian capitals: comparison between the year before and the first six first months of the COVID-19 pandemic. BMC Health Serv Res 21, 976 (2021). https://doi.org/10.1186/s12913-021-07006-x.
28. Janus SE, Makhlouf M, Chahine N, Motairek I, Al-Kindi SG. Examining Disparities and Excess Cardiovascular Mortality Before and During the COVID-19 Pandemic. Mayo Clin Proc. 2022 Dec;97(12):2206-2214. doi: 10.1016/j.mayocp.2022.07.008.
29. Jawad H Butt, Emil L Fosbøl, Thomas A Gerds, Charlotte Andersson, Kristian Kragholm, Tor Biering-Sørensen, et al. All-cause mortality and location of death in patients with established cardiovascular disease before, during, and after the COVID-19 lockdown: a Danish Nationwide Cohort Study, European Heart Journal, Volume 42, Issue 15, 14 April 2021, Pages 1516–1523, https://doi.org/10.1093/eurheartj/ehab028.
30. Coleman KM, Saleh M, Mountantonakis SE. Association between regional distributions of SARS-CoV-2 seroconversion and out-of-hospital sudden death during the first epidemic outbreak in New York. Heart Rhythm. 2021 Feb;18(2):215-218. doi: 10.1016/j.hrthm.2020.11.022.
31. Lange SJ, Ritchey MD, Goodman AB, Dias T, Twentyman E, Fuld J, et al. Potential Indirect Effects of the COVID-19 Pandemic on Use of Emergency Departments for Acute Life-Threatening Conditions - United States, January-May 2020. MMWR Morb Mortal Wkly Rep. 2020 Jun 26;69(25):795-800. doi: 10.15585/mmwr.mm6925e2.
32. Ohland J, Warren-Gash C, Blackburn R, Mølbak K, Valentiner-Branth P, Nielsen J, et al. Acute myocardial infarctions and stroke triggered by laboratory-confirmed respiratory infections in Denmark, 2010 to 2016. Euro Surveill. 2020;25(17):pii=1900199. https://doi.org/10.2807/1560-7917.ES.2020.25.17.1900199.
33. Francesca M, Raffaele C. Causal relationship between influenza infection and risk of acute myocardial infarction: pathophysiological hypothesis and clinical implications, European Heart Journal Supplements, Volume 22, Issue Supplement_E, June 2020, Pages E68–E72, https://doi.org/10.1093/eurheartj/suaa064.
34. Zou Z, Cini K, Dong B, Ma Y, Ma J, Burgner DP, et al. Time Trends in Cardiovascular Disease Mortality Across the BRICS: An Age-Period-Cohort Analysis of Key Nations With Emerging Economies Using the Global Burden of Disease Study 2017. Circulation. 2020 Mar 10;141(10):790-799. doi: 10.1161/CIRCULATIONAHA.119.042864.
35. Koskinas KC, Van Craenenbroeck EM, Antoniades C, Blüher M, Gorter TM, Hanssen H, et al. Obesity and cardiovascular disease: an ESC clinical consensus statement. Eur Heart J. 2024 Aug 30:ehae508. doi: 10.1093/eurheartj/ehae508.
36. Mafham MM, Spata E, Goldacre R, Gair D, Curnow P, Bray M, et al. COVID-19 pandemic and admission rates for and management of acute coronary syndromes in England. Lancet 2020; 396:381.

The Short Questionnaire to Assess Health-Enhancing Physical Activity in Syrian Adults' Populations

Mahfouz Al-Bachir[1], *Mohamad Adel Bakir*[2], *Husam Ahmad*[1]

(1) Department of Radiation Technology, Atomic Energy Commission of Syria
(2) Department of Nuclear medicine, Syrian Atomic Energy Commission, Syria

CORRESPONDING AUTHOR: *Dr. Mahfouz Al-Bachir, Department of Radiation Technology, Atomic Energy Commission of Syria, P.O. Box 6091, Damascus, Syria - Email: ascientific9@aec.org.sy*

SUMMARY

Background: To date, no studies have evaluated the reliability and/or validity of methods for measuring physical activity (PA) in free-living conditions within the Syrian population.
Methods: This study compared estimates of PA and sedentary behavior (SB) obtained from the ActiGraph WGT3X-TB (AG) accelerometer and the Short Questionnaire to Assess Health Enhancing Physical Activity (SQUASH). Forty-five adults (13 men and 32 women, mean age 36.9 ±8.3 years) completed the SQUASH twice, with a 45-day interval between administrations. Time spent in low, moderate, vigorous, and moderate-vigorous PA (MVPA) was calculated using the SQUASH and AG accelerometer data. Reliability was assessed by calculating the Spearman correlation coefficient between the PA items scores. Bland-Altman analysis was also performed. The validity of the SQUASH was determined using the AG accelerometer as the reference method.
Results: PA levels were systematically higher when measured by the SQUASH compared to the AG accelerometer. The Spearman's correlation coefficient for the overall SQUASH reproducibility was 0.64. The Spearman's correlation coefficient between the calculated total activity score from the SQUASH and the AG accelerometer was 0.31, indicating moderate reliability and validity of the SQUASH.
Conclusion: Given its simplicity, brevity, ease of use, and low cost, the SQUASH appears to be a suitable method for monitoring PA in Syrian adults. Further strengthening of the validity scores may be possible by providing more detailed information on the types of activities included in the questionnaire.

Keywords: Accelerometer, Self-report, Measurements, Reliability, Validity.

INTRODUCTION

Several urgent calls to action have been issued to address the global physical activity (PA) and sedentary behavior (SB) issues [1]. Enhancing PA among the population is crucial, as compelling evidence suggests that higher levels of PA are associated with better health outcomes [2]. Regular PA is essential to achieving and maintaining good health [3]. Low levels of PA are linked to an increased risk of morbidity and mortality [4,5]. Measuring PA can be challenging due to its diverse nature [6]. Therefore, an efficient and accurate tool for measuring PA in populations is essential. Physical activity is assessed using both objective measures like accelerometers and nonobjective measures based on self-report questionnaires [7]. At the population level, PA is often determined based on self-reported questionnaires [8,9]. Self-report questionnaires remain the most commonly used tool for assessing PA on a practical scale, although epidemiological studies often use questionnaires to determine PA levels [10]. The questionnaire is an inexpensive and useful method in categorizing subjects in low or high levels PA [11]. The Short Questionnaire to Assess Health Enhancing Physical Activity (SQUASH) is an example of such a questionnaire. SQUASH measures the activity score, a combination of intensity and duration of PA per week, and total minutes of activity per week [12]. The SQUASH is structured to facilitate comparisons with national and international PA recommendations [13].

DOI: 10.54103/2282-0930/25454
Accepted: 9th September 2024

The questionnaire includes questions about various activities such as commuting, gardening, odd jobs, sports, work activities, household chores, and leisure time. SQUASH is used by various research institutes and government agencies to monitor the PA behavior of youth and adult populations and adherence with PA guidelines [2,11,13].

Validation of PA measures is crucial to accurately assess and monitor the progress of PA interventions [14]. The accelerometer is often used as the gold standard method for assessing the validity of PA self-report questionnaires [15,16].

The purpose of this study was to determine the test reliability and validity of SQUASH in measuring self-reported habitual PA in a Syrian adult population using an ActiGraph WGT3X-BT accelerometer measurement device. While our previous work on using accelerometer to Assess Physical Activity Behavior in Syrian Adults in comparison with WHO recommendations reported that 1.5% of women and 6.7% of men, accumulate 150 minutes per week of MVPA with 10 minutes bouts [17]. Therefore, we expected that SQUASH would be relatively reliable in this population

MATERIAL AND METHODS

Study design, procedure and participants

Fifty-two adults (15 men and 37 women) were recruited from various workplaces within the Syrian Atomic Energy Commission (SAEC) in Damascus, Syria. Participants were selected according to the following inclusion criteria: (a) age range of 18-60 years; (b) willingness to wear accelerometers for seven days during free-living activities and sleeping; and (c) ability to complete surveys in Arabic. Participants signed written informed consent forms to participate in the study. The SAEC human ethics committee approved the study protocol, which was conducted in accordance with the Helsinki Declaration of the World Medical Association. The final sample consisted of 45 participants (13 men and 32 women) who met the inclusion criteria and had complete data on objectively assessed physical activity (PA), height, and weight. No technical errors were encountered during accelerometer registrations. Participants were required to wear the accelerometer for at least four days with a minimum of 600 minutes of valid daily monitor wear. The study involved three visits to the participants' workplaces. During the first visit, trained research assistants conducted measurements of anthropometric parameters. Body weight was measured using electronic scales (Seca, Model: 7671321004; Germany) with an accuracy confirmed by using known masses (20 kg). Height was measured to the nearest 0.5 cm using a wall-mounted stadiometer (Seca, Model: 225 1721009; Germany). Body mass index (BMI) was calculated as weight (kg) divided by height (m) squared. A demographic questionnaire was administered to collect information on sex, age, marital status, education levels, and smoking status. Participants were asked to wear the ActiGraph accelerometer on their left hip with an adjustable elastic belt for seven days. Seven days after the first visit, participants returned the accelerometers and completed the SQUASH-1 questionnaire. The third visit occurred 45 days later, during which the SQUASH-2 questionnaire was completed.

Accelerometer processing

The material requirements for the study included the ActiGraph (AG) (WGT3X-BT, Pensacola. FL. 32502 USA) accelerometer, a small, lightweight, tri-axial activity monitor that provides data on physical activity (PA), including activity counts, energy expenditure (kcal), steps, and activity intensity (METs) [18]. Participants wore a single AG accelerometer unit over their left hip, attached to an elasticized waistband, for all waking hours during a 24-hour period over seven days. Participants were instructed to remove the device before engaging in aquatic activities such as swimming, bathing, and showering. The AG accelerometer data were processed using the Actilife software and exported to Microsoft Excel format. Within Microsoft Excel, mean minutes of PA, including light, moderate, and vigorous, as well as sedentary behavior (SB) were calculated per day. The daily average was then multiplied by seven to create a weekly total activity score [19].

Short Questionnaire to Assess Health Enhancing Physical Activity (SQUASH)

The SQUASH was selected to evaluate the physical activity (PA) behavior of the study population.. For activities at work and household, the intensity was pre-defined into two categories: light or intense [20]. The total minutes of activity were calculated for each question by multiplying frequency (days/week) by duration (min/day) while the missing answers were not considered the farther calculation. The total minutes of PA per week were calculated by aggregating the total minutes of PA reported in the SQUASH.

To assess the reproducibility of the SQUASH over time, 45 subjects completed the questionnaire in a randomised order on two separate occasions, with a 45-day interval between the first and second measurements. This time period was chosen to avoid recall bias while preventing significant changes in PA levels. The validity of the self-reported PA questionnaire was assessed by correlating the SQUASH total scores with the AG accelerometer results [12,21].

Diagnosis Criteria

Cutoff points for intensity categories were based on the PA guideline [21]. Based on the reported effort in the SQUASH activities were given an intensity score (ranging from 1 to 9). For example; walking and odd jobs activities received an intensity score of 1, 2, or 3 based on reported effort being low, moderate, or vigorous activity, respectively. For gardening and bicycling these intensity scores were 4, 5, and 6, respectively. The scores for light and intense activity at work and household are 2 and 5, respectively [21].

Statistical analysis

Participants with both SQUASH and AG accelerometer data were included in the present analysis. All statistical analyses were performed using the Statistical Package for Social Science (SPSS) for Windows (Version 17.0.1, 2001, SPSS Inc., Chicago, USA). Continuous variables were expressed as mean ± standard deviation (SD), whereas categorical variables were represented by frequency and percentage. Statistical significance was set at $p < 0.05$ or corresponding p-value in all tests.

Spearman's rho correlations were used to evaluate relationships between change scores of self-reported (SQUASH) and objective measurement (AG accelerometer) of PA. To examine differences between pre and post levels of PA, Wilcoxon signed rank tests were used. Effect sizes (r) of the differences between the baseline, an effect size of r = 0.0 to 0.30 was considered a negligible effect, 0.30 to 0.50 small effect, 0.50 to 0.70 moderate effect, 0.70 to 0.90 high effect, and 0.90 to 1.00 a large effect [21]. Summary data is reported as mean (SD) and statistical significance was assumed at $p < 0.05$.

RESULTS

Baseline characteristics of study participants

Demographic characteristics of the study populations are presented in Table 1. Of the 45 individuals who participated in the study, (n=32; 71.1%) were female and (n=13; 28.9%) were male. The mean age of the participants in the study group was 36.9±8.3 years, with 42.6±5.3 years for men and 34.6±8.2 years for women. The mean BMI was 26.6±4.7 kg/m^2, with 28.2±3.3 kg/m^2 for men and 26.0±5.1 kg/m2 for women. The majority of the total sample of the participants were aged 30–45 years (n = 27; 60.0%), married (n = 30; 66.7%); overweight/obese (n = 28; 62.2%), with high school level education (secondary school or higher) (n = 38; 84.5%); and non-smokers (n = 23; 51.1%) (Table 1).

Table 1. Descriptive demographic characteristics of the study participants.

Variables	Subcategory	Total sample	Men	Women
		N=45	N=13 (28.9%)	N=32 (71.1%)
Age (years)	Mean (±SD)	36.9±8.3	42.6±5.3	34.6±8.2
Age group (n, %)*	18-29	10 (22.2)	0 (0.0)	10 (31.3)
	30-45	27 (60.0)	8 (61.5)	19 (59.4)
	>45	8 (17.8)	5 (38.5)	3 (9.4)
Marital status (n, %)	Single	15 (33.3)	2 (15.4)	13 (40.6)
	Married	30 (66.7)	11 (84.6)	19 (59.4)
BMI (Kg/m^2)	Mean (±SD)	26.6±4.7	28.2±3.3	26.0±5.1
BMI Category (n, %)*	Normal Weight	17 (37.8)	2 (15.4)	15 (46.9)
	Overweight/Obese	28 (62.2)	11 (84.6)	17 (53.1)
Educational level (n, %)*	< Secondary School	7 (15.6)	3 (23.1)	4 (12.5)
	Secondary School	8 (17.8)	5 (38.5)	3 (9.4)
	> Secondary School	30 (66.7)	5 (38.5)	25 (78.1)
Smoking status (n, %)*	Yes	22 (48.9)	10 (76.9)	12 (37.5)
	No	23 (51.1)	3 (23.1)	20 (62.5)

** Significant deference exists between men and women at p<0.05*

Measurements of PA and Test–retest reliability

All descriptive statistics for SQUASH-1 and SQUASH-2 are presented in Table 2. Of the reported time (SQUASH-1), 14% was spent during leisure-time activities (walking, cycling, gardening, odd jobs and sports), 60% during household activities (light and intense) and 25% at work (light and intense). Almost no time (less than 1.5%) was spent on commuting activities (walking and cycling) (Table 2).

The Spearman's correlation coefficient for the total activity score was 0.64. For the other, separate questions the Spearman's correlation coefficient ranged from - 0.08 to 0.84. Reliability for commuting bicycling activities could not be measured. Intense commuting by walking was least reliable (r=-0.08), while, light activities at work were most reliable (r=0.84) (Table 2).

Bland and Altman analysis for the total activity score showed significant differences between the two SQUASH assessments, with the most observations staying at the 0 – 1.96 SD range and within the 95% limits of agreement (Figure 1). The limits of agreement for estimated total activities that derived from SQUASH-1 vs SQUASH-2. MVPA that derived from SQUASH, with upper and lower 95% LOA of -5624.5 to 5677.7 min/week at 2 SD (Figure 1).

Significant (p < 0.074) inter-method agreement was demonstrated between SQUASH-1 and SQUASH-2 estimates. Correlations between the SQUASH-1 with SQUASH-2 were R2 = 0.004, and p=0.651.

Validity of the SQUASH

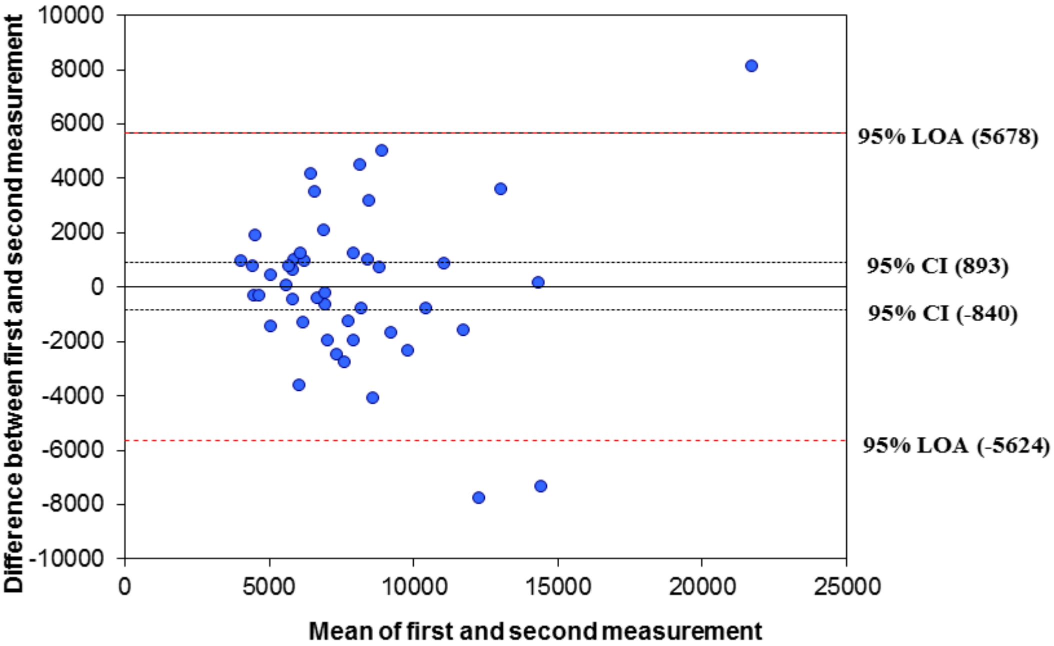
Assessment of physical activity (PA) using the SQUASH revealed significantly higher total PA minutes across all intensity levels compared to the AG accelerometer. Moderate and vigorous intensity PA durations were consistently higher when assessed by the SQUASH than by the AG accelerometer (Table 3). The SQUASH reported an average of 3409 ± 1102 minutes of total weekly activity, while the AG accelerometer recorded 1395 ± 436 minutes of total weekly activity. Predominantly, both the SQUASH (78%) and the AG accelerometer (83%) indicated that most of the time was spent in low-intensity activities. The Spearman's correlation coefficient between the total activity scores derived from the SQUASH and AG accelerometer was 0.31 (95% CI ranging from 1690 to 2336, as shown in Table 3).

Table 2. Test–retest reliability of SQUASH-SF based on Spearman-rank correlation coefficients (n = 45)

Item	Minutes/week	Activity score	Activity score	Reliability
	SQUASH-1	SQUASH-1	SQUASH-2	R Spearman
	n= 45	n = 45	n = 45	n = 45
All items together	3415 (1095)	7973 (3516)	8000 (3699)	0.64**
Commuting				
Walking	46 (54)	100 (125)	113 (176)	0.35*
Cycling	0 (0)	0 (0)	0 (0)	-
Leisure time				
Walking	163 (251)	342 (523)	450 (649)	-0.08
Cycling	1 (9)	7 (45)	23 (157)	-0.02
Gardening	63 (325)	313 (1623)	255 (1149)	0.83**
Odd jobs	207 (666)	340 (1030)	193 (704)	0.24
Sports	29 (75)	112 (335)	205 (682)	0.60**
Activities at work				
Light	661 (801)	1321 (1601)	1584 (2004)	0.84**
Intense	214 (337)	1071 (1687)	808 (1121)	0.26
Household activities				
Light	1937 (407)	3873 (814)	4040 (833)	0.19
Intense	101 (466)	506 (2327)	167 (917)	0.65**

*Minutes per week spent in different categories of physical activity (mean ± SD), activity scores from the dual measurements (mean ± SD) and reliability of the total activity scores, as well as reliability of the scores on separate questions of the SQUASH (Spearman correlation coefficient). *p<0.05, **p<0.01.*

Figure 1. Bland and Altman graph with the limits of agreement (LOA).



The difference between total activity scores on the first and second SQUASH plotted against their mean for each patient, together with 95% confidence interval (CI) and the 95% LOA.

Activity score = minutes x intensity.

Table 3. Physical activity levels based on SQUASH -SF and ActiGraph results (n = 45)

	SQUASH-1 (n=45)	ActiGraph (n=45)	d	SE d	95% CI	rSpear-man
Total	3409 ± 1102	1395 ± 436	2013 ± 1076	160	1690 - 2336	0.31*
Low intensity	2667 ± 920	1155 ± 372	1511 ± 955	142	1224 - 1798	0.01
Moderate intensity	388 ± 573	235 ± 106	153 ± 526	78	-5 - 311	0.46**
Vigorous intensity	354 ± 550	5 ± 6	349 ± 551	82	183 - 514	-0.03

All times are expressed as minutes of activity per week (mean ± SD).

d = mean difference between time spent in physical activity as assessed by the first administered SQUASH (SQUASH-1) and the Actigraph.

SE d = standard error of the mean difference. 95% CI = 95% confidence interval of the mean difference between the two measurements.

However, the Bland and Altman analysis demonstrated poor agreement between the two methods in calculating total PA, with an R value of 0.741 and p-value of 0.000. The 95% limits of agreement (LOA) ranged from -94.8 to 4121.3 minutes per week at 2 standard deviations (Figure 2).

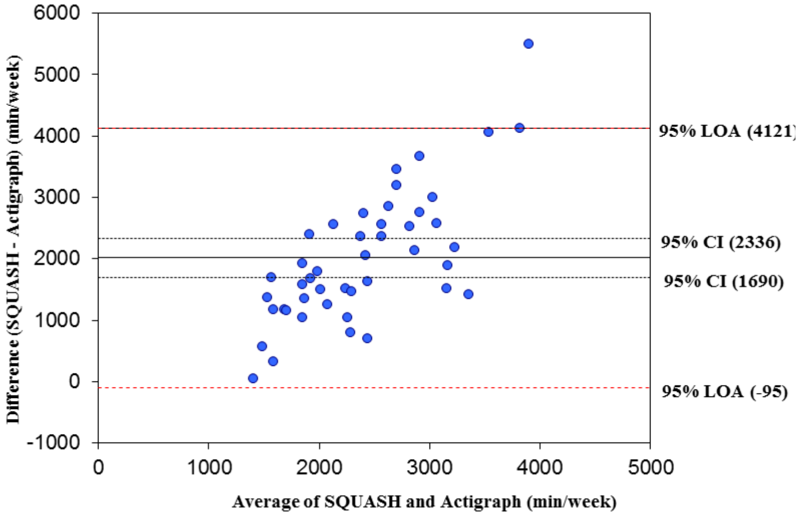
Figure 3 illustrates the distribution of weekly minutes of moderate-equivalent PA from self-reported data and AG accelerometer measurements.

Both distributions show a wide variation in physical activity levels within the sample, with the majority exhibiting low activity levels and a smaller proportion engaging in higher levels of activity. The AG accelerometer data showed a broader range of values compared to self-reported activity. Ideally, a bell-shaped curve would represent a normally distributed dataset. When comparing tertiles of activity scores with activity counts, the exact agreement was 48%, and the weighted kappa value was 0.20 (Figure 3).

DISCUSSION

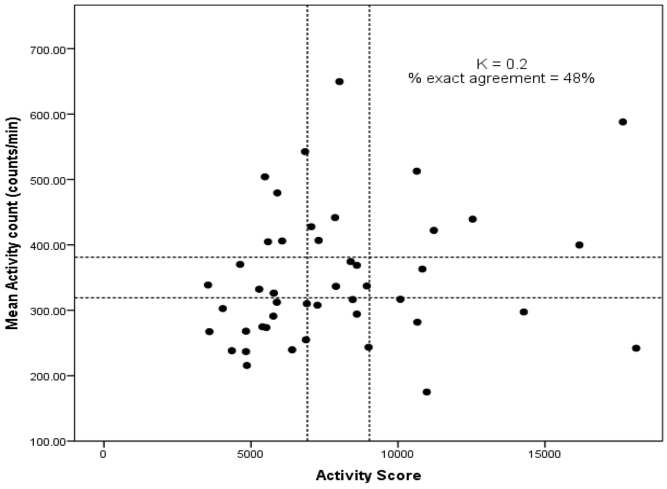
The primary objective of this study was to assess the reliability and validity of the SQUASH questionnaire in measuring physical activity (PA) and sedentary behavior (SB) among Syrian adults aged 18-60 years, using the AG accelerometer as the comparison method. The AG accelerometer is considered one of the most reliable tools for comparing subjective methods like the SQUASH [23,24]. We evaluated the measurement properties of the SQUASH as a more detailed self-report method for assessing PA behavior. Questionnaires have been recommended as a reliable and valid approach for assessing PA levels in the general adult population [11,13]. However, the SQUASH cut-off points appear to be higher, leading to an overestimation of PA levels compared to the AG accelerometer. This finding supports the well-

Figure 2. Bland and Altman graph with the limits of agreement (LOA).



The difference between total minutes of physical activity per week as assessed by mean of the SQUAH and the Actigraph, plotted against their mean for each patient, together with 95% confidence interval (CI) and the 95% LOA.

Figure 3. Tertiles (dotted lines) of the activity score per week (SQUAH) and the mean activity counts per minutes (Actigraph).



documented observation that the SQUASH and other PA questionnaires tend to overestimate PA levels [25-27].

Reliability of the SQUASH

The Spearman correlation for overall (all items together) reliability of the SQUASH in our assessment was 0.64. Therefore, the reproducibility is comparable to other PA questionnaires. The correlation coefficient we determined in this validation experiment of SQUASH is within the range of correlation coefficient indicating in the literatures [11,28,29]. Therefore, the correlation coefficient in our study can be considered as reasonable and acceptable. However, the reproducibility of PA questionnaires has been determined in adults in many countries in the past. Most of the researchers found that kappa values varying from 0.47-0.89 [11,28,29]. This result may be explained by the real differences in the PA levels from week to week, overestimation of the real PA attribute to an incorrect perception of activity, misunderstanding of the questions that led to measurement error in both questionnaire surveys, and/or the inability to correctly recall all activities performed when completing the surveys [27,30].

Validity of the SQUASH

The main finding of the present study is that the agreement between self-reported physical activity (PA) using the SQUASH questionnaire and objectively assessed PA using the ActiGraph (AG) accelerometer was small (r=0.31). To the best of our knowledge, this is the first study to examine the validity of the SQUASH in a Syrian population using AG accelerometer measurements. Therefore, we are unable to directly compare our data to others in similar populations.

Previous studies have reported Spearman's correlation coefficients ranging from 0.31 to 0.45 based on the validity of other PA questionnaires [25,27]. The SQUASH total scores were reported to be somewhat better, with correlations of 0.45 in healthy adults and 0.35 in patients [11,21]. In a review assessing the validity of seven PA questionnaires in adults, the correlation with accelerometer total counts ranged from 0.34 to 0.89 [13,31]. However, it has been suggested that the correlation between total PA assessed by a questionnaire and total accelerometer counts should be at least 0.50 to demonstrate construct validity [32]. Based on these guidelines, the SQUASH did not reach the standard for construct validity in Syrian adults.

Our finding is consistent with the recommendation by Cleland et al. who suggested that self-reported physical activity in general was not sufficiently accurate for individual assessment. These study results have important implications for how PA is being assessed and reported among the population in Syria. The present findings suggest that the SQUASH may not be appropriate for assessing minutes of PA in Syrian adult populations.

Clinical Implications and Future Directions

These findings have potential important clinical implications, as PA measurement using self-reports has been suggested for use in healthcare settings for risk assessment [33,34]. To improve the validity of questionnaires in adult populations, it may be possible to strengthen the validity scores by providing more detailed examples of the types of activities adults may do [35,36]. This may enhance their ability to accurately recall their activities during the 7-day reporting period.

Some potential sources of information errors could arise in PA records. Despite these limitations, the SQUASH is a suitable tool for monitoring and assessing PA behavior in Syrian adults and could provide useful PA information for international comparisons. The results showed moderate reliability and validity of the questionnaire, which is in agreement with data published in review papers [9,37-39].

Strengths of the study

The strength of our study is that we used objectively measured PA data (AG accelerometer), and were able to compare it with subjective data (SQUASH). To our best knowledge, this study is the first to validate the SQUASH to the Arabic language. We detected moderate correlation between PA estimated by AG accelerometer and SQUASH. As well as a moderate test-retest reliability for most of the SQUASH items. In this study we used the AG accelerometer as reference method, thereby enhancing the criterion validity of the research.

Limitations of the study

This study has several limitations that should be considered when interpreting the results.

Small Sample Size: The study population was relatively small, which may limit the generalizability of the findings.

Healthy Adult Participants: Only healthy adults were recruited for the study. Therefore, the results cannot be extrapolated to other populations such as young people or older adults.

Lack of Gold Standard Comparison: Estimates of sedentary behaviour (SB) and physical activity (PA) were not compared with a gold standard method like doubly labeled water, which is considered the most accurate technique for measuring energy expenditure.

Limited Generalisability: The generalisation of the results is limited to a similar study population. Comparisons of PA and SB using these methods with other study samples could provide inaccurate results.

Higher Education Level: Although the researchers tried to cover a large socioeconomic range, the participants had a higher than average education level compared to the general Syrian population. Physical activity in people from higher socioeconomic groups is more accurately recalled [40,41].

Variation Between Questionnaires: Several participants varied in their responses between questionnaires. Some questions may be regarded as easier or more difficult to respond to, and this should be taken into consideration when testing the validity of a question or questionnaire.

Reliability Affected by Interview Interval: Previous research has shown that reliability can be affected by the interval between the two interviews [42].

In conclusion, while this study provides valuable insights into the physical activity and sedentary behavior patterns of healthy Syrian adults, the limitations mentioned above should be considered when interpreting the findings and designing future research in this area.

CONCLUSION

The current study marks the first validation of an international standardized physical activity (PA) questionnaire for Syrian adults. The study aimed to assess the reliability and validity of the SQUASH in monitoring PA levels. The results indicate that the SQUASH has moderate reliability and validity in determining minutes of PA. Despite this, the SQUASH was found to be a suitable method for monitoring PA in Syrian adults. Future validation studies should consider using doubly labeled water as a criterion for employment in surveillance.

CONFLICT OF INTEREST

The authors declare no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGMENTS

The authors wish to express their deep appreciation to the Director General of AECS. This study was supported by the International Atomic Energy Agency under the Technical Research Contract No. SYR/6/012 is gratefully acknowledged.

REFERENCES

1. Cleland C, Ferguson S, Ellis G, Hunter RF. Validity of the International Physical Activity Questionnaire (IPAQ) for assessing moderate-to-vigorous physical activity and sedentary behavior of older adults in the United Kingdom. BMC Medical Research Methodology 2018;18:176.
2. Campbell N, Gaston A, Gray C, Rush E, Maddison R, Prapavessis H. The short questionnaire to assess health enhancing (SQUASH) physical activity in adolescents: a validation using doubly labeled water. Journal of Physical Activity and Health 2016;13:154-158.
3. Nicaise V, Crespo NC, Marshall S. Agreement between the IPAQ and accelerometer for detecting intervention related changes in physical activity in sample of Latin women. Journal of Physical Activity and Health, 2014;11:846-852.
4. Doherty A, Jackson D, Hammerla N, PloÈtz T, Olivier P, Granat MH, et al. Large-Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLoS ONE, 2017;12(2):e0169649.
5. Carlin A, Perchoux C, Puggina A, Aleksovska K, Buck C, Burns C, et al. A life course examination of the physical environmental determinants of physical activity behaviour: A Determinants of Diet and Physical Activity (DEDIPAC) umbrella systematic literature review. PLoS ONE, (2017;12(8):e0182083.
6. Janz K. Physical activity in epidemiology: Moving from questionnaire to objective measurement. British Journal of Sports Medicine, 2006;40(3):191–192.
7. Tomioka K, Iwamoto J, Saeki K, Okamoto N. Reliability and Validity of the International Physical Activity Questionnaire (IPAQ) in Elderly Adults: The Fujiwara-kyo Study J Epidemiol., 2011;21(6):459-65.
8. Strath SJ, Kaminsky LA, Ainaworth BE, Ekelund U, Freedson PS, Gary RA., et al. Guide to assessment of physical activity: clinical and research applications: a scientific statement from the American Heart Association. Circulation, 2013;128:2259-2279.
9. Helmerhorst HJF, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. The International Journal of Behavioral Nutrition and Physical Activity, 2012;9:103.
10. Sarkin JA, Nichols JF, Sallis JF, Calfas K.J. Self report measures and scoring protocols affect prevalence estimates of meeting physical activity guidelines. Med Sci. Sports Exerc., 2000;32(1):149-156.
11. Wendel-Vos GCW, Schuit AJ, Saris WHM, Kromhout D. Reproducibility and relative validity of the short questionnaire to assess health enhancing physical activity. Journal of Clinical Epidemiology. 2003;56:116-1169.
12. Seves, BL, Hoekstra F, Schoenmakers JWA, Brandenbarg P, Hoekstra T, Hettinga FJ, et al. Test-retest reliability and concurrent validity of the Adapted Short Questionnaire to Assess Health-enhancing physical activity (Adapted-SQUASH) in adults with disabilities. JOURNAL OF SPORTS SCIENCES. 2020;

https://doi.org/10.1080/02640414.2020.1850983. ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjsp20.

13. Wagenmakers R, Akker-Scheek I, Groothoff JW, Zijlstra W, Bulstra SK, Kootstra JWJ, Wendel-Vos GCW, Raaij JJAM, Stevens M. Reliability and validity of the short questionnaire to assess health-enhancing physical activity (SQUASH) in patients after total hip arthroplasty. BMC Musculoskeletal Disorders. 2008;9:141. http://www.biomedcentral.com/1471-2474/9/141.
14. Hidding LM, Chinapaw MJM, van Poppel MNM, Mokkink LB, Altenburg TM. An updated systematic review of childhood physical activity questionnaires. Sports Med.2018;48:2797–2842.
15. Hagströmer M, Trost SG, Sjöström M, Berrigan D. Levels and patterns of objectively assessed physical activity - a comparison between Sweden and the United States. American Journal of Epidemiology. 2010:171:1055-1064.
16. Loney T, Standage M, Thompson D, Sebire SJ, Cumming S. Self-Report vs. Objectively Assessed Physical Activity: Which Is Right for Public Health? J Phys Act Health. 2011;8:62–70.
17. Al-Bachir M, Ahmad H. Using accelerometer Analysis to Assess Physical Activity and Sedentary Behavior in Syrian Adults. Epidemiology Biostatistics and Public Health. 2023; 18(1).
18. Yano S, Koohsari MJ, Shibata A, Ishii K, Mavoa S, Oka K. Assessing Physical Activity and Sedentary Behavior under Free-Living Conditions: Comparison of Active Style Pro HJA-350IT and ActiGraphTM GT3X+. International Journal of Environmental Research and Public Health. 2019;16:3065.
19. Oyeyemi AL, Umar M, Oguche F, Aliyu SU, Oyeyemi AY. Accelerometer-Determined Physical Activity and Its Comparison with the International Physical Activity Questionnaire in a Sample of Nigerian Adults. PLoS ONE. 2014;9(1): e87233.
20. Nicolaou M, Gademan MGJ, Snijder MB, Engelbert RHH, Dijkshoorn H, Terwee CB, Stronks K. Validation of the SQUASH physical activity questionnaire in a multi-ethnic population: the HELIUS study. PLoS ONE. 2016;11(8):1-14.
21. Arends S, Hofman M, Kamsma YPT, van der Veer E, Houtman PM, Kallenberg C, Spoorenberg A, Brouwer E. Daily physical activity in ankylosing spondylitis: validity and reliability of the IPAQ and SQUASH and the relation with clinical assessments. Arthritis Research and Therapy, 2013;15(4):R 99.
22. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi Med J., 2012;24(3):69-71.
23. Lorenzo B, Donatella C, Angela Polito. "Assessment of Physical Activity in a Group of Adults in Italy: Comparison of Two Different Methodologies." Journal of Physical Activity Research, 2017;2(2):117-123.
24. Sirard JR, Pate RR. "Physical activity assessment in children and adolescents". Sports medicine, 2001;31(6):439-454.
25. Wolin KY, Heil DP, Askew S, Matthews CE, Bennett GG. Validation of the International Physical Activity Questionnaire-Short among Blacks. J Phys Act Health, 2008;5(5):746–60.
26. Bauman A, Ainsworth BE, Bull F, Craig CL, Hagstromer M, Sallis JF, et al. Progress and pitfalls in the use of the International Physical Activity Questionnaire (IPAQ) for adult physical activity surveillance. J Phys Act Health, 2009; 6 Suppl 1, S5–8.
27. Medina C, Barquera S, Ian J. Validity and reliability of the International Physical Activity Ques¬tionnaire among adults in Mexico. Rev Panam Salud Publica. 2013;34(1):21–8.
28. Pols MA, Peeters PH, Ocke MC, Bueno-de-Mesquita HB, Slimani N, Kemper HC, Collette HJ. Relative validity and repeatability of new questionnaire on physical activity. Prev Med., 1997;26(1):37-43.
29. Roeykens J, Rogers R, Meeusen R, Magnus L, Borns J, de Meirleir K. Validity and reliability in Flemish population of WHO- MONICA Optional Study of Physical Activity Questionnaire. Med Sci. Sports. Exerc., 1998;30(7):1071-1075.
30. Shephard RJ. Limits to the measurement of habitual physical activity by questionnaires. Br J Sports Med., 2003;37(3):197–206.
31. Sallis JF, Saelens BE. Assessment of physical activity by self report: status, limitations, and future directions. Res Q Exerc Sport, 2000;71:s1-14.
32. Terwee CB, Mokkink LB, van Poppel MN, Chinapaw MJ, van Mechelen W, de Vet HC. Qualitative attributes and measurement properties of physical activity questionnaires: a checklist. Sports Med, 2010;40:525-537.
33. 33. Ekblom Ö, Ekblom-Bak E, Bolam KA, Ekblom B, Schmidt C, Söderberg S, Bergström G, Börjesson M. Concurrent and predictive validity of physical activity measurement items commonly used in clinical settings– data from SCAPIS pilot study. BMC Public Health, 2015;15:978. http://dx.doi.org/10.1186/s12889-015-2316-y/
34. Vanhees L, Geladas N, Hansen D, Kouidi E, Niebauer J, Reiner Z, et al. Importance of characteristics and modalities of physical activity and exercise in the management of cardiovascular health in individuals with cardiovascular disease (Part III). European Journal of Preventive Cardiology, 2012;19(6):1333–1356. doi:10.1177/2047487312437063.
35. Clemes SA, David BM, Zhao Y, Han X, Brown WJ. Validity of two self-report measures of sitting time. J Phys Act Health, 2012;9:533–539.
36. Atkin AJ, Gorely T, Clemes SA, Yates T, Edwardson C, Brage S, Salmon J, Marshall SJ, Biddle SJ. Methods of measurement in epidemiology: sedentary behaviour. Int J Epidemiol, 2012:41:1460–1471.
37. Lee PH, Macfarlane DJ, Lam TH, Stewart SM. Validity of the International Physical Activity Questionnaire Short Form (IPAQ-SF): a systematic review Int. J Behav Nutr Phys Act. 2011; 8:115. Doi:10.1186/1479-5868-8-115.
38. van Poppel MN, Chinapaw MJ, Mokkink LB, van Mechelen W, Terwee CB. Physical activity questionnaires for adults: asystematic review of measurement properties. Sports medicine, 2010;40(7):565-600. Doi:10.2165/11531930-000000.
39. Baumeister SE, Ricci C, eKohler S, Fischer B, Topfer C, Finger JD, Leitzmann MF. Physical activity surveil-

lance in the European Union: reliability and validity of the European Health interview Survey-Physical Activity Queastionnaire (EHIS-PAQ). International Journal of Behavioral Nutrition and Physical Activity, 2016;13:61. Doi 10.1186/s12966-016-0386-6.

40. Conklin AI, Forouhi, NG, Suhrcke, M, Surtees P, Wareham NJ, Monsivais P. Socioeconomic status, financial hardship and measured obesity in older adults: a cross-sectional study of the EPIC-Norfolk cohort. BMC Public Health, 2013;13:1039.

41. Innerd P, Harrison R, Coulson M. Using open source accelerometer analysis to assess physical activity and sedentary behaviour in overweigh and obese adults. BMC Public Health, 2018;18:543. https://doi.org/10.1186/s12889-018-5215-1

42. Tran DV, Lee AH, Au TB, Nguyen CT, Hoang DV. Reliability and validity of the international physical activity questionnaire short form for older adults in Vietnam. Health Promot J Austr., 2013:24(2):126-31. doi: 10.1071/HE13012.

High Prevalence of Metabolic Syndrome among Female Vegetable Market Traders in Hargeisa, Somaliland: Risk Factors and Implications

Fosia A. Mohamoud[(1)], *Arthur Kwena*[(1)], *Caleb Nyamwange*[(1)], *Geoffrey K. Maiyoh*[(1)]

(1) Department of Biochemistry and Clinical Chemistry, School of Medicine, Moi University, P.O Box 4606-30100, Eldoret, Kenya.

CORRESPONDING AUTHOR: Geoffrey K. Maiyoh, Department of Biochemistry and Clinical Chemistry, School of Medicine, Moi University, P.O Box 4606-30100, Eldoret, Kenya. E-mail: kattam@mu.ac.ke

SUMMARY

Background: Metabolic syndrome, characterized by abdominal obesity and two or more of the following components (fasting blood glucose ≥100 mg/dL, low HDL-cholesterol, high triglycerides, and hypertension), is a common cause of morbidity and mortality. In Somaliland, female vegetable market vendors, who often sit for long hours, face an elevated risk.

Aims: This study aims to assess the prevalence and associated factors of metabolic syndrome in this population.

Methods: Conducted from December 2020 to April 2021 in Hargeisa's vegetable markets, this cross-sectional study recruited 291 women using stratified convenience random sampling. Structured questionnaires collected socio-demographic data, while fasting blood samples provided information on blood sugar, triglycerides, and high-density lipoprotein levels. Descriptive statistics and logistic regression were used for analysis.

Results: A total of 291 women, aged 21-80 years (mean age 45.3 (12.3) years, participated. The prevalence of metabolic syndrome was 71.8%, significantly higher than global averages. High waist circumference (87.9%, P = 0.00) was the most prevalent component, suggesting unique dietary or lifestyle factors. Notably, no significant association was found between marital status and metabolic syndrome (P = 0.41), contrasting with findings from other regions. Approximately 45% of participants had two components of metabolic syndrome, 40% had three components, and 15% had four components, respectively, indicating a distinct pattern of component distribution.

Conclusions: This study found a high prevalence of metabolic syndrome (71.8%) in this population. Key risk factors included older age, high BMI, and increased waist-to-hip ratio, highlighting the need for targeted health interventions and education for this specific occupational group.

Keywords: Metabolic syndrome; Abdominal obesity; Prevalence; Risk factors; Hargeisa, Somaliland.

ABBREVIATIONS

AACE	American Association of Clinical Endocrinology	**HDL**	High-Density Lipoprotein
BMI	Body Mass Index	**IDF**	International Diabetes Federation
CHD	Coronary heart disease	**MS**	Metabolic Syndrome
CI	Confidence Interval	**MTRH**	Moi Teaching and Referral Hospital
CVD	Cardio Vascular Disease	**NCEP-ATP III**	National Cholesterol Education Odds ratio Program Adult Treatment Panel III
EGIR	European Group for the study of Insulin Resistance	**OR**	Odds ratio
FBS	Fasting Blood Sugar	**WHO**	World Health Organization
		WHR	Waist to Hip Ration

DOI: 10.54103/2282-0930/24218
Accepted: 1st October 2024

INTRODUCTION

Metabolic syndrome (MS) is a complex condition that has reached epidemic proportions globally, posing a significant public health challenge. It is characterized by the concurrent presence of at least three out of five cardiometabolic abnormalities: obesity, hyperglycemia, hypertriglyceridemia, reduced high-density lipoprotein (HDL) levels, and hypertension [1, 2]. These factors significantly elevate the risk of cardiovascular diseases and type 2 diabetes, underscoring the syndrome's public health relevance. The development of MS is influenced by various risk factors, including obesity, aging, and prolonged sedentary work [3].

Prompt detection of MS is vital for initiating lifestyle interventions that can mitigate the risk of related chronic conditions. Diagnostic criteria have been developed by numerous health organizations, such as the European Group for the Study of Insulin Resistance (EGIR), the National Cholesterol Education Program Adult Treatment Panel III (NCEP: ATP III), the World Health Organization (WHO), the American Association of Clinical Endocrinology (AACE), and the International Diabetes Federation (IDF) [4, 5]. Despite these differing standards, they all emphasize the necessity of managing MS to avert its severe health impacts.

Globally, the prevalence of MS varies from 12.5% to 31.4%, depending on the diagnostic criteria used [6]. Regional disparities are evident; for example, prevalence rates among women in Middle Eastern countries range from 11.7% in Kuwait to 41% in Saudi Arabia. In contrast, African nations report rates from 17.9% in Ethiopia to 40.2% among Kenyan women [7, 8]. In Central Africa, Bowo-Ngandji et al. (2023) [9], documented a similarly high prevalence of MS among women, particularly those exposed to urbanized environments and sedentary occupations. These findings underscore the urgent need for region-specific interventions tailored to the occupational and lifestyle factors prevalent in African settings.

Recent data from Africa suggest that MS is becoming increasingly prevalent. A comprehensive systematic review by Whelton et al. (2018) [10] found the overall prevalence of MS in African populations to be 32.4% (95% CI: 30.2–34.7), based on 297 studies from 29 African countries involving 156,464 participants. The prevalence was higher among women (36.9%) compared to men (26.7%) and was significantly elevated in adults over the age of 18 (33.1%) compared to children under 18 years (13.3%) [10]. Moreover, the prevalence of MS was particularly high among individuals with type 2 diabetes (66.9%) and those with cardiovascular diseases (48.3%). Despite these alarming statistics, Charles-Davies et al. (2023) [11] argue that existing diagnostic criteria may not fully account for the genetic and environmental factors unique to African populations, advocating for the development of African-specific diagnostic cut-offs to enhance the accuracy of diagnosis and management.

In Hargeisa, Somaliland, a substantial proportion of women work in vegetable markets. In the present study, five of the main markets in Hargeisa, with an approximate population of 1,000 female vegetable vendors, were sampled. These women face unique occupational challenges that may heighten their risk of MS. In particular, the market traders endure long hours of sedentary work, limited access to healthy dietary options, and exposure to stress, all of which are recognized risk factors for MS [3]. Understanding these occupational and lifestyle factors is essential for designing targeted interventions aimed at reducing the burden of MS in this vulnerable population.

While data on MS in Somaliland is limited, its prevalence is increasing across African nations, including neighboring countries [10]. Studies show that women are particularly affected, with MS rates ranging from 17.9% in Ethiopia to 40.2% in Kenya [7, 8]. However, there has been little research on MS among female vegetable market traders in Hargeisa. This study aims to address this gap by examining the prevalence and risk factors of MS in this population. The findings are essential for guiding public health strategies and interventions to mitigate MS's impact on women in Somaliland and similar contexts.

MATERIALS AND METHODS

Study Design and Participants

This cross-sectional study was conducted between December 2020 and April 2021 at vegetable markets in Hargeisa, the capital city of Somaliland, which covers an area of 78 km^2. Hargeisa has an estimated total female population of 500,000, of which approximately 300,000 to 325, 000 are women aged over 18 years [12].

The study employed stratified convenience random sampling by selecting five main vegetable markets in Hargeisa, which account for a significant portion of the overall market volume and sales. This approach enhances the representativeness of the vendor demographics. Additionally, we considered accessibility for data collection, diversity of offerings, and practical constraints such as time and resources.

All adult women (>18 years) working in the vegetable markets during the study period were eligible for enrollment, except those who were pregnant.

Sample size determination

The minimum sample size for the study was determined by the use of Fisher's formula for sample size calculation using the prevalence of 24% obtained from a study by Tran et al. [1], from neighboring Ethiopia.

n = Z2pq/d2

Where:

n = Desired sample size (population >10,000), population greater than 10,000

Z = the standard normal deviate usually set at 1.96 which corresponds to 95% confidence level.

p = Estimated characteristic of the study population 24.0% prevalence of metabolic syndrome among women in Ethiopia [1].

q = 1 – p

d = the minimum error/degree of accuracy desired, which is usually set at 5% or 0.05

Therefore:

(1.96)2 *0.24*0.76/0.0025

The initially determined sample size for the study was 280 participants. However, to account for potential non-responses and ensure sufficient statistical power, a total of 309 participants were selected proportionately across five markets. This adjustment was made to accommodate any non-responders and minimize the risk of an inadequate sample size. A total of 309 potential participants were initially interviewed, with 18 respondents not meeting eligibility criteria resulting in a final enrollment rate of 94.4%. Consequently, the final analysis is based on the responses from 291 participants.

Sampling Procedure

The sampling procedure for this study was designed to ensure proportional representation of vegetable sellers across five main markets in Hargeisa. A total of 1,000 vegetable sellers were estimated to be working in these markets, with a sample size of 309 sellers determined for the study [13]. The following steps were taken to distribute the sample across the markets:

Market 1: Out of 143 vegetable sellers, 45 were sampled, representing 14.5% of the total population.

Market 2: This market had the largest population of sellers, with 312 vendors. A sample of 96 sellers was taken, accounting for 30.1% of the total sample size.

Market 3: From the 255 vendors in this market, 78 sellers were included in the study, representing 25.2% of the total sample.

Market 4: In this market, 44 sellers were selected from a total of 140 vendors, which accounted for 14.2% of the total population.

Market 5: Lastly, 46 sellers were sampled from 150 total vendors, contributing 15% to the overall sample.

This stratified random sampling ensured proportional representation across the markets, with the total sample size of 309 sellers reflecting 30.9% of the total vegetable vendor population in the five main markets of Hargeisa.

Data Collection

Demographic Data

The questionnaire used for data collection was meticulously developed through a collaborative process involving experts in socio-economic and demographic research [14]. This ensured the inclusion of comprehensive and relevant questions tailored to the study's objectives. The structured questionnaire was designed to capture socio-economic and demographic information, including age (in years), marital status (e.g., single, married, divorced, widowed), education level (e.g., no formal education, primary, secondary, tertiary), duration of working at the vegetable market (in years), and family history of chronic illnesses (e.g., diabetes, hypertension).

On the first day of data collection, the interviewer-administered questionnaire was employed to gather the necessary information. For participants who faced challenges with reading and writing, the questionnaire was explained to ensure accurate and complete responses. To further enhance the validity of the data, the questionnaire was reviewed and revised with participants. All interviews were conducted in a secluded room to maintain privacy and confidentiality. Anthropometric measurements (weight, height, and abdominal circumference) were measured using standardized techniques and calibrated equipment, as outlined by Utkualp (2015) [15] and WHO, (2011) [16]. BMI was calculated by dividing weight by height squared (kg/m^2) and classified according to WHO criteria (≥30 kg/m^2). Waist and hip circumferences were measured with participants standing, using a Roche circumference tape. The waist-to-hip ratio (WHR) was calculated by dividing waist circumference by hip circumference, both measured in centimeters. All measurements were taken by trained nurses to ensure accuracy and consistency.

Clinical and Laboratory Data

Blood pressure measurements were obtained using an Omron digital sphygmomanometer (Kyoto, Japan) after at least 10 minutes of rest following the 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults [10]. To ensure consistency in the fasting period, participants received standardized written instructions for a 12-hour overnight fast. Participants were organized into manageable groups with staggered fasting start times to facilitate efficient blood collection. Blood samples were drawn by qualified phlebotomists over several days, with 28 to 30 participants per day. Venous blood samples (5 ml) were collected via venipuncture and placed in red-topped tubes for biochemical analysis of fasting blood sugar (FBS) and lipid profiles. FBS was measured using a glucometer (Acon Diabetes Care, Pennsylvania, USA), while triglycerides and HDL levels were analyzed using an automated chemistry analyzer (Mindray BA-88A, India). This systematic approach, incorporating staggered fasting, scheduled blood draws, and qualified personnel, ensured both efficiency and accuracy across all participants.

Ensuring Quality and Reliability in Biochemical Measurements

The quality and reliability of biochemical measurements were maintained through several quality control measures. First, standardized equipment was used for all tests, including a glucometer (Acon Diabetes Care, Pennsylvania, USA) for fasting blood sugar (FBS) and an automated chemistry analyzer (Mindray BA-88A, India) for triglycerides and HDL levels. These devices were regularly calibrated according to manufacturer guidelines to ensure accuracy. Additionally, all blood samples were collected following a strict 12-hour fasting period, reducing variability in the test results. Laboratory personnel were trained to follow standard operating procedures (SOPs) for handling and analyzing samples, further ensuring consistency. Daily internal quality control tests were also conducted on the equipment to detect and correct any potential deviations before running participant samples.

Definition of Metabolic Syndrome

The International Diabetes Federation criteria for metabolic syndrome (MS) were used in this study [17]. The rationale for choosing the International Diabetes Federation (IDF) criteria over others was due to its global applicability, ease of use in clinical practice, and its emphasis on central obesity as a key factor in metabolic syndrome [18]. These criteria are widely accepted, particularly in regions with a high prevalence of central obesity and diabetes such as in many regions of Africa and Asia [19]. According to these criteria, an abdominal circumference greater than 80 cm is essential for diagnosing MS, along with any two of the following components:

a. Fasting glucose ≥100 mg/dL or ≥6.1 mmol/L
b. Triglycerides ≥150 mg/dL
c. High-density lipoprotein cholesterol ≤50 mg/dL
d. Hypertension ≥130 mmHg systolic or ≥85 mmHg diastolic

Statistical Analysis

The data in this study were analyzed using Epi Info software. Continuous variables were summarized using descriptive statistics, specifically the mean and standard deviation. Associations between metabolic syndrome and specific variables, including age, Body Mass Index (BMI), and waist-to-hip ratio, were identified by calculating Odds Ratios (OR) and p-values using logistic regression. A p-value of <0.05 was considered statistically significant for hypothesis testing. Incomplete records were omitted during final data analysis.

RESULTS

Socio demographic Characteristics of the Participants

A total of 291 women were studied. The mean age of the participants was 45.32 years, with the youngest being 21 and the oldest 80 (Table 1). The largest proportions of the participants, approximately 32.7%, were in the 31-40 age groups. The mean BMI was 27.14 (±6.67) kg/m^2, and about 31.6% of the women were classified as obese (Table 1). High waist-to-hip ratio (WHR) was

observed in 69.8% of the participants (Table 1). A marked proportion of women had been working for more than 10 years. Regarding marital status, 63.9% of the women were married, and 18.6% were separated (Table 1).

Prevalence of Metabolic Syndrome

The prevalence of metabolic syndrome among women was 71.8% (Table 2). The prevalence for individual components of metabolic syndrome were as follows: high waist circumference (87.9%), low high-density lipoprotein cholesterol levels (78.3%), high blood pressure (56.1%), high fasting glucose levels (51.2%), and high triglyceride levels (42.6%) in the studied population (Table 2). Among the five metabolic syndrome components, high waist circumference (87.9%) and low HDL levels (78.3%) were the most prevalent. All components were highly associated with metabolic syndrome (p=0.000) except HDL (p=0.1667) (Table 2).

Table 1: Socio-demographic and Clinical Characteristics of the Participants

Variable	Frequency (%)	Mean (SD)
Age groups in years		45.32 (12.3)
21-30	32 (11.0)	
31-40	95 (32.6)	
41-50	84 (28.9)	
51-60	49 (16.8)	
61-70	27 (9.3)	
71-80	4 (1.4)	
Marital status		
Married	186 (63.9)	
Single	24 (8.3)	
Widow	27 (9.3)	
Separated	54 (18.6)	
Working duration		6.04 (5.55)
0-5 years	157 (54.9)	
6-10 years	97 (34.3)	
More than 10 years	37 (12.7)	
BMI		27.14 (6.67)
Obesity	92 (31.7)	
Overweight	92 (31.6)	
Normal weight	81 (27.8)	
Underweight	26 (8.9)	
WHR (Waist-to-Hip Ratio)		0.90 (0.11)
High WHR (>0.85)	203 (69.8)	
Normal WHR (0.81-0.85)	53 (18.2)	
Low WHR (<0.81)	35 (12.0)	

Table 2: Association between Metabolic Syndrome Components and Risk Factors with Odds Ratios

Variable (IDF)	Mean (SD)	Total Frequency (n, %)	With MS (%)	95% CI	Odds Ratio	p-value
Participants	-	291 (100)	71.8%	-	-	-
MS		209 (71.8%)	100%			
Normal		82 (28.2%)	0%			
Waist Circumference	97.91(15.30)			1.0474 - 1.0936	1.07	<0.001
Normal		35 (12.1%)	-			
High		256 (87.9%)	81.64%			
Triglycerides	151.53 (94.62)			1.0101 - 1.0206	1.02	<0.001
Normal		167 (57.4%)	-			
High		124 (42.6%)	91.94%			
HDL Cholesterol	34.55 (27.30)			0.9851 - 1.0026	0.99	0.167
Normal		63 (21.7%)	-			
Low		228 (78.3%)	76.75%			
Fasting Blood Sugar	6.30 (2.97)			1.3674 - 2.5540	1.87	<0.001

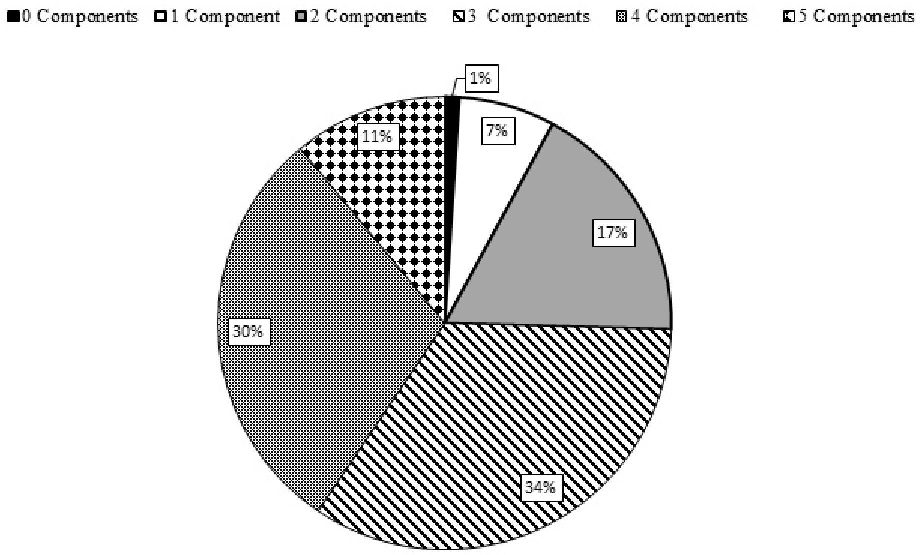
Normal		142 (48.8%)	-			
High		149 (51.2%)	91.28%			

Blood Pressure	Sys
	133.7(23.3)
	Dis
	80.8(14.8)
				3.465 - 10.936	6.16	<0.001

Normal		128 (43.9%)	-			
High		163 (56.1%)	87.12%			

As shown in Figure 1, the largest proportion of study participants (34%) had three metabolic syndrome components, followed by those with four (30%), two (17%), five (11%), and none (1%).

Figure 1: The prevalence rates MS individual components (high waist circumference, low high-density lipoprotein cholesterol levels, high blood pressure, high fasting glucose levels and high triglyceride level) among participants were determined by dividing the number of women with each specific MS component by the total number of participants in the study and multiplying by 100.



Factors Associated with Metabolic Syndrome

The prevalence of metabolic syndrome varied across different age groups: 59.4% in the 21-30 age group, 62.1% in the 31-40 age group, 79.7% in the 41-50 age group, 75.6% in the 51-60 age group, 85.2% in the 61-70 age group, and 100% in the 71-80 age group (Table 3). Approximately 74.3% of participants who had been working for 6-10 years had metabolic syndrome (Table 3). Among obese participants, 83.6% had metabolic syndrome, while 82.6% of overweight participants were diagnosed with metabolic syndrome. A summary of factors associated with metabolic syndrome among women working in the vegetable markets in Hargeisa is provided in Table 3.

Table 3: Factors Associated with Metabolic Syndrome among Women Working in the Vegetable Markets in Hargeisa

Variable	Diagnosed with MS Frequency (%)	Without MS Frequency (%)	p-Value
Age groups (years)			
21-30 (N = 32)	19 (59.4)	13 (40.6)	≤0.000c
31-40 (N = 95)	59 (62.1)	36 (37.9)	
41-50 (N = 84)	67 (79.7)	17 (20.3)	
51-60 (N = 49)	37 (75.6)	12 (24.4)	
61-70 (N = 27)	23 (85.2)	4 (14.8)	
71-80 (N = 4)	4 (100)	0	
Marital status			0.4151
Single (N = 24)	16 (66.7)	8 (33.3)	
Married (N = 186)	132 (70.9)	54 (29.1)	
Divorced (N = 54)	41 (75.9)	13 (24.1)	
Widow (N = 27)	20 (74.1)	7 (25.9)	
Working duration			0.3385
0-5 years (N = 157)	109 (69.4)	48 (30.6)	
6-10 years (N = 97)	72 (74.3)	25 (25.7)	
More than 10 years (N = 37)	28 (75.6)	9 (24.4)	
BMI categories			≤0.000c
Obesity (> 30 kg/m^2)	77 (83.6)	15 (16.4)	
Overweight (25.0-29.9 kg/m^2)	76 (82.6)	16 (17.4)	
Normal weight (18.5-24.9 kg/m^2)	0	81 (100)	
Underweight (< 18.0 kg/m^2)	0	26 (100)	
WHR categories			≤0.000c
High WHR (> 0.85 cm)	159 (78.3)	44 (21.7)	
Normal WHR 0.18-0.85 cm		57(100)	
Low WHR < 0.80 cm		31(100)	

WHR: Waist to hip ration; BMI: body mass index; c: Refers to values that are highly significant (p ≤ 0.001).

DISCUSSION

Our study investigated the prevalence and contributing factors of metabolic syndrome (MS) among female vendors at vegetable markets in Hargeisa, Somaliland. With a prevalence rate of 71.8%, this study highlights a serious public health concern in this population. Although the vegetable selling profession involves prolonged periods of sitting, we acknowledge that directly linking this sedentary work to MS warrants further investigation. Nonetheless, this finding raises concerns about potential health risks associated with prolonged sedentary behavior, which has been noted in other studies to increase cardio-metabolic risk [20].

The sociodemographic characteristics of the participants revealed that most were middle-aged, with a mean age of 45.32 years, a factor known to be associated with an increased risk of MS [21]. Moreover, the participants exhibited an average BMI of 27.14 kg/m^2, with 31.7% classified as obese, reinforcing the global trend that links higher BMI with MS [22]. The study also found that high waist circumference and low HDL-cholesterol levels were prevalent among participants, aligning with findings in other populations where abdominal obesity and low HDL are closely linked to MS, particularly in women [23, 24].

In our analysis, we considered dietary habits and lifestyle factors as potential contributors to the high MS prevalence. While dietary data were not directly assessed, existing literature suggests that diets high in refined carbohydrates and unhealthy fats may contribute to the development of MS [25]. Future studies should prioritize dietary and lifestyle assessments to better understand their role in this population.

Although prolonged sitting is a plausible risk factor for MS among vegetable vendors, this study did not directly measure lifestyle behaviors such as physical activity levels. Structural or social factors, such as the nature of their work environment and cultural expectations, may limit these women's ability to be physically active during or after work. This highlights the need for future research to investigate the specific environmental and social determinants that may contribute to sedentary behavior and MS in this population [26].

Our findings are consistent with studies conducted in other countries, such as Ethiopia, where a prevalence of 22% was reported among working adults [1]. The differences in prevalence between these regions may be attributable to variations in socioeconomic status, occupational structures, healthcare access, and cultural factors, all of which require further exploration to contextualize these disparities. In South Africa, for instance, occupational diversity and differences in health system accessibility might partly explain why our findings in Somaliland present a higher prevalence rate of MS [4, 5]. Although the cross-sectional nature of our study precludes causal inferences, the high prevalence of MS observed in this population underscores the need for targeted interventions. These interventions should focus on promoting physical activity and dietary improvements while addressing the specific challenges faced by this occupational group. Future research should focus on longitudinal studies to assess the progression of MS, as well as qualitative studies to understand the barriers to adopting healthier lifestyles.

CONCLUSION

This study revealed a high prevalence of metabolic syndrome (71.8%) among female vegetable market traders in Hargeisa, Somaliland. Key factors associated with this condition included older age, elevated BMI, and increased waist-to-hip ratio, raising concerns about potential health risks related to sedentary behavior. While this study does not establish causality, the findings suggest the need for targeted interventions and health education tailored to this population. Future research should further explore the occupational and lifestyle factors contributing to metabolic syndrome in this group to guide the development of effective strategies for improving their health outcomes.

AUTHOR CONTRIBUTIONS

FAM, AK, CN, and GKM did conception and planning of the study. FAM ran the laboratory assays and abstracted patient demographic and clinical data. FAM did data entry and cleaning. FAM and CN conducted the data analysis. FAM and GKM did the drafting of the manuscript. FAM, AK, WEI, CN and GKM reviewed the manuscript for philosophical insights. All the authors reviewed the final manuscript and approved it for submission.

FUNDING

The authors did not receive support from any organization for the submitted work.

ETHICAL APPROVAL

Verbal informed consent was obtained from all participants. The study was approved by the Institutional Research and Ethics Commission at Moi University/MTRH (REF IREC/2017/148) and the Directorate of Health Services and Hospitals, Ministry of Health Development in Somaliland.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author, [G.K.M], upon reasonable request.

CONFLICT OF INTERESTS

The authors declared that they have no conflict of interest.

ACKNOWLEDGEMENT

The authors extend their gratitude to Moi University and Hargeisa Hospital for the provision of essential facilities that facilitated this study. Special thanks are also due to Yasmin Tahir and Ibrahim Gulled for their indispensable contributions to the data collection process.

REFERENCES

1. Tran A, Gelaye B, Girma B, Lemma S, Berhane Y, Bekele T, et al. Prevalence of metabolic syndrome among working adults in Ethiopia. International journal of hypertension. 2011;2011.
2. Setayeshgar S, Whiting SJ, Vatanparast H. Metabolic syndrome in Canadian adults and adolescents: prevalence and associated dietary intake. International Scholarly Research Notices. 2012;2012.
3. Rochlani Y, Pothineni NV, Kovelamudi S, Mehta JL. Metabolic syndrome: pathophysiology, management, and modulation by natural compounds. Therapeutic advances in cardiovascular disease. 2017;11(8):215-25.
4. Motala AA, Mbanya J-C, Ramaiya KL. Metabolic syndrome in sub-Saharan Africa. Ethnicity & disease. 2009;19:8-10.
5. Kaduka LU, Kombe Y, Kenya E, Kuria E, Bore JK, Bukania ZN, et al. Prevalence of metabolic syndrome among an urban population in Kenya. Diabetes care. 2012;35(4):887-93.
6. Noubiap JJ, Nansseu JR, Lontchi-Yimagou E, Nkeck JR, Nyaga UF, Ngouo AT, et al. Geographic distribution of metabolic syndrome and its components in the general adult population: A meta-analysis of global data from 28 million individuals. Diabetes research and clinical practice. 2022;188:109924.
7. Okafor CI. The metabolic syndrome in Africa: Current trends. Indian journal of endocrinology and metabolism. 2012;16(1):56-66.
8. Ansarimoghaddam A, Adineh HA, Zareban I, Iranpour S, HosseinZadeh A, Kh F. Prevalence of metabolic syndrome in Middle-East countries: Meta-analysis of cross-sectional studies. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2018;12(2):195-201.
9. Bowo-Ngandji A, Kenmoe S, Ebogo-Belobo JT, Kenfack-Momo R, Takuissu GR, Kengne-Nde C, et al. Prevalence of the metabolic syndrome in African populations: A systematic review and meta-analysis. PloS one. 2023;18(7):e0289155.
10. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Journal of the American College of Cardiology. 2018;71(19):e127-e248.
11. Charles-Davies MA, Ajayi OO. Redefining the Metabolic Syndrome in Africa: A Systematic Review between 2005 and 2022. Dubai Diabetes and Endocrinology Journal. 2023;29(2):89-98.
12. Development SMoPaN. Somaliland Demographic and Health Surveys. Hargesisa: Government of Somaliland; 2020.
13. Kotrlik J, Higgins C. Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research. Information technology, learning, and performance journal. 2001;19(1):43.
14. DeVellis RF, Thorpe CT. Scale development: Theory and applications: Sage publications; 2021.
15. Utkualp N, Ercan I. Anthropometric measurements usage in medical sciences. BioMed research international. 2015;2015(1):404261.
16. Organization WH. Waist circumference and waist-hip ratio: report of a WHO expert consultation, Geneva, 8-11 December 2008. 2011.
17. Alberti KGMM, Zimmet P, Shaw J. Metabolic syndrome—a new world- wide definition. A consensus statement from the international diabetes federation. Diabetic medicine. 2006;23(5):469-80.
18. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, et al. Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity. Circulation. 2009;120(16):1640-5.
19. Misra A, Khurana L. Obesity and the metabolic syndrome in developing countries. The Journal of Clinical Endocrinology & Metabolism. 2008;93(11_supplement_1):s9-s30.
20. Dunstan DW, Healy GN, Sugiyama T, Owen N. Too much sitting and metabolic risk—has modern technology caught up with us. European Endocrinology. 2010;6(1):19-23.
21. Ford ES, Giles WH, Mokdad AH. Increasing prevalence of the metabolic syndrome among US adults. Diabetes care. 2004;27(10):2444-9.
22. Mottillo S, Filion KB, Genest J, Joseph L, Pilote L, Poirier P, et al. The metabolic syndrome and cardi-

ovascular risk: a systematic review and meta-analysis. Journal of the American College of Cardiology. 2010;56(14):1113-32.

23. Grundy SM, Brewer Jr HB, Cleeman JI, Smith Jr SC, Lenfant C. Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. Circulation. 2004;109(3):433-8.
24. Assmann G, Gotto Jr AM. HDL cholesterol and protective factors in atherosclerosis. Circulation. 2004;109(23_suppl_1):III-8-III-14.
25. Esposito K, Chiodini P, Colao A, Lenzi A, Giugliano D. Metabolic syndrome and risk of cancer: a systematic review and meta-analysis. Diabetes care. 2012;35(11):2402-11.
26. Proper KI, Singh AS, Van Mechelen W, Chinapaw MJ. Sedentary behaviors and health outcomes among adults: a systematic review of prospective studies. American journal of preventive medicine. 2011;40(2):174-82.

A Meta-Analysis Approach on Medical, Surgical and Expectant Management on Abortion of First Trimester

Shivali Negi(1), *Kavya Sharma*(1), *Anwesa Acharya*(2), *Ananya Prabhu*(3), *Rinshu Dwivedi*(4), *Ramesh Athe*(5)

(1) Master Student, Centre for Public Health (U.I.E.A.S.T), Panjab University, Chandigarh, India-160014.
(2) Master Student, Department of CSE, CMR University, Karnataka, India-562149.
(3) MDS Student, Manipal College of Dental Sciences, Manipal, Karnataka, India-575001.
(4) Assistant Professor, National Institute of Technology, Hamirpur, Himachal Pradesh, India-177005.
(5) Assistant Professor, Indian Institute of Information Technology, Dharwad, Karnataka, India-580029.

CORRESPONDING AUTHOR: Dr Ramesh Athe, Assistant Professor, Indian Institute of Information Technology, Dharwad, Karnataka, India-580029. Email: dr.athe9@gmail.com

SUMMARY

An increase in miscarriage in the first trimester of gestation and its associated complication is burden-some on the quality of life of a woman. Medical, surgical, and expectant care are carried out after the miscarriage to remove any remaining tissues in the uterus. Understanding the efficacy and safety of these interventions will raise awareness and be a deciding factor to choose an appropriate treatment plan. Present review aims to determine the efficacy and safety of medical, surgical, and expectant care of various medical and surgical methods for first-trimester miscarriage. This review included studies that allocated women to medical, surgical or expectant management in the first trimester. PubMed, Cochrane Library, MEDLINE, and Embase Library were searched for the literature. The primary outcome was the complete evacuation of products of conception. Data were independently reviewed, graded for evidence quality, and assessed for risk bias by using the guidelines of PRISMA (Preferred Report Items for Systematic Review and Meta-Analysis). 21 eligible articles were included in this systematic review, comprising of 7931 patients undergoing medical, surgical or expectant-management for early spontaneous-miscarriage. The success rate in surgical intervention was higher when compared with medical intervention (OR: 16.12 [9.11, 28.52]) and expectant management (OR: 2.78 [2.13, 3.61]). Whereas medical intervention had a high success rate when compared with expectant-management (OR: 4.29 [2.31, 7.97]). The review determines the effect of medical, surgical, and expectant-management procedures on women who have had spontaneous-miscarriages in their first-trimester. PROSPERO-International prospective register of systematic reviews–CRD42020154395.

Keywords: Surgery; Medical; Expectant management; Spontaneous Abortion; First trimester; Systematic review.

INTRODUCTION

A miscarriage is a common occurrence defined as a nonviable pregnancy with an empty/incomplete gestational sac, an embryo without cardiac action, or a gestational trophoblastic illness with molar placental degradation. It occurs in 15% to 20% of pregnancies, according to estimates. Approximately 80% of these spontaneous miscarriage pregnancies occur between the first and thirteenth weeks of gestation, with the risk decreasing after 12 weeks. Most patients are unaware of how frequently spontaneous miscarriages occur in the first trimester, which can lead to anxiety (30%), post-traumatic stress disorder (34%), and sadness (10%), all of which can disrupt mental harmony [1-3].

As a preventive measure for the evacuation of the retained products of conception in missed miscarriage and incomplete miscarriage, therapeutic alternatives

DOI: 10.54103/2282-0930/22705
Accepted: 15th July 2024

such as surgical evacuation, expectant management, and medicinal management are used[4]. Vacuum aspiration is a type of surgical uterine evacuation that involves a vacuum source. It is also known as suction curettage, endometrial aspiration, or mini-suction. It is possible to utilize a handheld vacuum syringe or mechanical pump that is operated by foot (Manual Vacuum Aspiration) or electricity (Electric Vacuum Aspiration)[5]. Sharp metal curettage (also known as dilatation and curettage) is commonly performed in an operating room while the patient is sedated or under a general or regional anesthetic[6].

Miscarriage medications typically involve synthetic prostaglandins such as Misoprostol, which is used primarily in incomplete miscarriages. Mifepristone, a progesterone antagonist, is used in conjunction with misoprostol to treat early miscarriage, particularly missed/silent miscarriage. Misoprostol, a safe and cheap medication, may allow for early POC ejection while avoiding complications[7,8]. The approach of expectant management allows the retained tissues of gestation to usually pass naturally, outside the hospital, and is an alternative to standard treatment with medication or surgery[9].

Surgical procedure has a 95% success rate for missed abortion but an important unresolved issue is the cost of surgery and the risks associated with anesthesia[5]. Medical management of miscarriages has been demonstrated to be advantageous, particularly in women who have had a missed miscarriage or an empty sac. Misoprostol, on the other hand, is not approved for usage in all countries[10]. If a miscarriage is not handled, the fetal tissue will normally pass naturally, as it did for more than 65% of women who suffered a miscarriage with may take up to two weeks. Unexpected hospitalizations and surgical curettage, on the other hand, occurred significantly more frequently during expectant and medicinal management than following surgical management[5,11].

The main aim of this systematic review is to determine the efficacy and safety of medical, surgical, and expectant care of different medical and surgical methods for first-trimester miscarriage.

METHODOLOGY

The systematic review and meta-analysis were performed interpretation to the PRISMA and registered in Prospero CRD42020154395[12-14]. The PICO strategy (population, intervention, comparison, and outcome) was used to build the research question. Thus, this systematic review is required to clarify the safety, efficacy, and side effect of medical, surgical, and expectant management on first-trimester spontaneous miscarriage.

Eligibility

The review included original articles that evaluated the safety, efficacy, and side effect of pharmacological, surgical and expectant management on first-trimester spontaneous miscarriage. Studies that patients did not receive medical, surgical and expectant interventions for miscarriage, review articles, letters to the editor; in vitro studies, conference articles and case reports or series were excluded from the present study[15].

Search strategy

A literature search on Medline/PubMed, Cochrane Library, MEDLINE, and Embase Library was performed using mesh terms mentioned in **Supplementary material S1** and were searched[14,15]. Randomized case-control, cohort studies, and quasi-trials of women with first-trimester miscarriage were included, and directed a systematic review and meta-analysis generated both direct and mixed evidence on the effectiveness and side effects of medical, surgical, and expectant management. The selected articles through these databases were de-duplicated and the titles and abstracts of the articles were read independently by two of the authors using the software Rayyan. The studies which could potentially cover the inclusion criteria for this review were identified at this stage and accessed in their entirety. Cases of disagreement were resolved by consensus.

Data Extraction

Randomized trials, quasi-randomized studies, cohort study and case-control studies that evaluated medical treatment, surgical treatment and expectant treatment management of first-trimester miscarriage that was defined as a spontaneous loss of a non-viable intrauterine pregnancy between 0 and 13th weeks gestation were included. Studies that evaluated combination of two treatment options (e.g. medical, expectant and surgical management) were included. Studies with multiple comparison arms were also included. We manually extracted data, using a excel sheet on: year and author, country of study, sample size, age, confounding factors, type of intervention, pre-outcomes and outcomes: success rate, bleeding, abdominal pain, and infection rate[14,15].

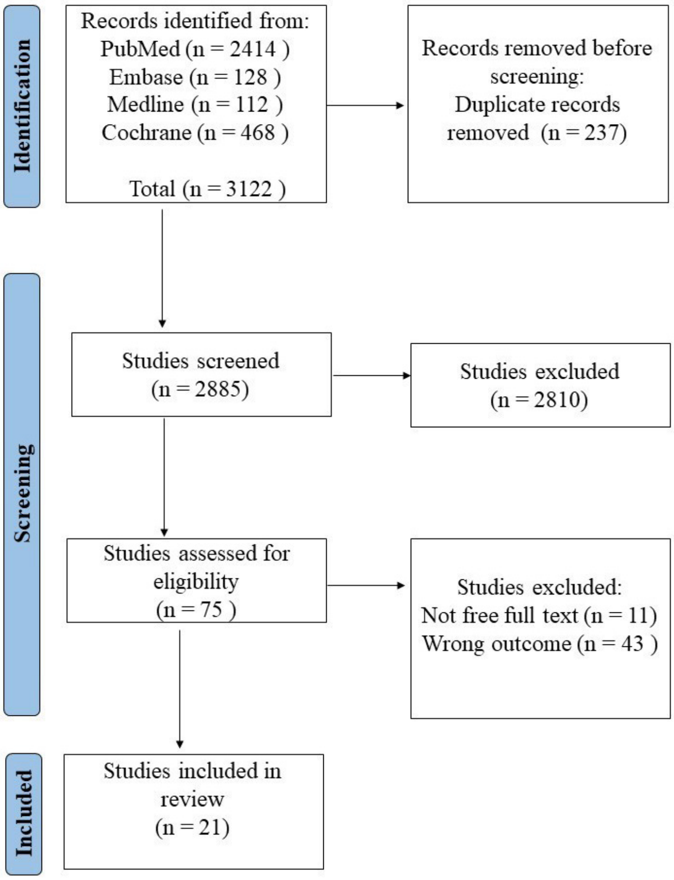
Assessment of risk of bias in included studies

The risk of bias for the chosen studies was evaluated with Joanna Briggs Institute (JBI) criteria[16]. Two reviewers independently will decide whether there is a "High risk", "Low risk" or "unclear risk" of bias. The risk of bias will be ranked high when the study reached up to 49% of yes, moderate when it is (50-69) % and low when it is above or equal to 70%.

2.5 Statistical Analysis

The meta-analyses were performed for suitable outcomes using Review Manager Software 5.4.1. The odds Ratio (OR) was used as an effective measure for dichotomous variable outcomes in the study such as success rate, surgery required abdominal pain, blood diffusion, infection rate, nausea, and vaginal bleeding. The weighted mean difference was used for vaginal bleeding in days. The heterogeneity between the medical, surgical, and expectant studies was verified by the inconsistency test (I^2). I^2 values lower than 25% were considered low heterogeneity among the studies; values between 25 and 49% were considered moderate heterogeneity and values greater than 50% were considered high heterogeneity. When I^2 was equal to 0 the fixed effects model was used, when I^2 was greater than 0 the random effects model was used. The dependent variable was success rate, vaginal bleeding, abdominal pain and infection rate [14,15,17-21]. Statistical analyses were performed with Review Manager (RevMan) software version 5.4.1, and Comprehensive Meta-Analysis (CMA) software trial version (www.meta-analysis.com).

Figure 1. Represents the PRISMA flowchart for study selection


Identification
Records identified from:
PubMed (n = 2414)
Embase (n = 128)
Medline (n = 112)
Cochrane (n = 468)
Total (n = 3122)
Records removed before screening:
Duplicate records removed (n = 237)
Screening
Studies screened (n = 2885)
Studies excluded (n = 2810)
Studies assessed for eligibility (n = 75)
Studies excluded:
Not free full text (n = 11)
Wrong outcome (n = 43)
Included
Studies included in review (n = 21)

RESULTS

3122 articles were identified from the literature, 2414 in PubMed, 112 in Medline, 128 in Embase and 468 in Cochrane. 237 studies were duplicate studies in the databases and were excluded from the study. After full screening of articles based on inclusion and exclusion criteria there were 21 eligible articles were included in this systematic review, comprising of 7931 patients undergoing medical, surgical or expectant management for early spontaneous miscarriage[22-42] and depicted in **Figure 1** and also and summary statistics tabulated in **Table 1.** Represents the PRISMA flowchart for study selection.

Study characteristic

Eleven studies compared medical intervention with surgical [23,27,31,33-40], three studies compared medical management with expectant management[26,32,39], and 8 studies compared surgical with expectant management [22,24,28,30,31,39-41]. Out of the 21 articles included, sixteen had randomized controlled trial design [22,24-39,41,42], two had quasi controlled design [28,40] and three were cohort studies [23,36,37]. The primary demographic characteristics of all the included 21 studies are tabulated**.** Complete abortion was defined as complete expulsion of the products of conception without any additional management. We could compare the success rate of the intervention, and for the reported side effects, we could only compare the incidence of abdominal pain, vaginal bleeding and infection.

Risk of bias assessment

The risk of bias was estimated using the JBI scale; most studies showed low to moderate risk of bias. The lowest risk of bias was seen in study by Demetroulis[25] et al., and highest risk of bias was seen among Fernlund[26] et al. Most studies did not conduct statistical analysis for confounding factors. Blinding of participants and clinicians was not possible due to the type of intervention. The results of the quality assessment of the studies are shown in the **Supplementary Table S2.**

Meta-analysis

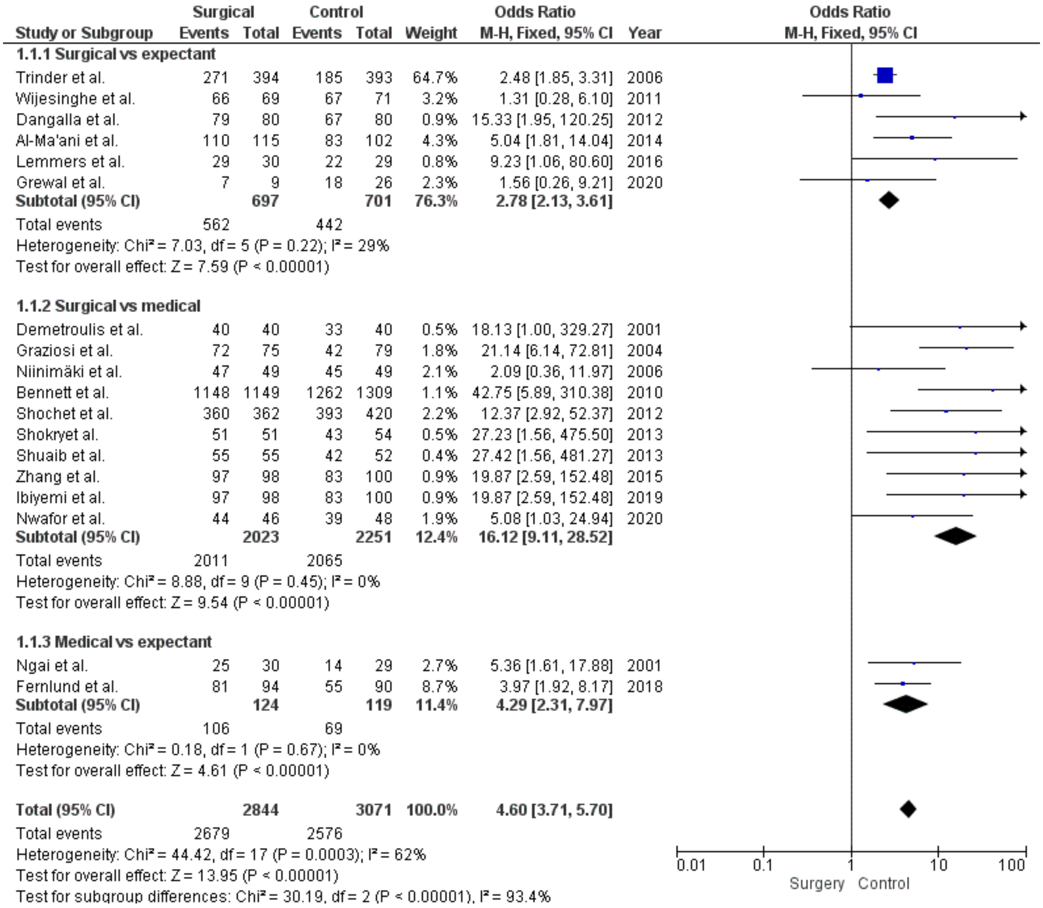
The results of meta-analysis for the outcomes are presented as forest plots in **Figure 2**. The forest plot indicated that the odds of success in surgical intervention was higher when compared with medical intervention (N= 4274, OR: 16.12 [9.11, 28.52], Heterogeneity: $Chi^2 = 7.03$, $df = 5$ ($P = 0.22$); $I^2 = 29\%$) and expectant management (N=1398, OR: 2.78 [2.13, 3.61], Heterogeneity: $Chi^2 = 7.03$, df =

Table 1: Summary of the trials assessing the characteristics abortion

S.No.	Year	Authors	Location	Study design	Study Duration	Intervention/ Control	Mean age	Gestation mean	Parity
1	2014	Al-Ma'ani et al.22	Germany	RCT	30	Expectant vs surgical	32.5	62.5 days	N/A
2	2010	Bennett et al.23	US	Cohort	3	Medical, MVA	2.5	N/A	N/A
3	2012	Dangalla et al.24	Sri Lanka	RCT	14	Expectant care vs ERPC	29	9.2	52 (64.6)
4	2001	Demetroulis et al.25	UK	RCT	10	Misoprostol and D&C	28.4	72.8	12
5	2018	Fernlund et al.26	Sweden	RCT	30	Misoprostol vs expectant	32.2	76.5	45
6	2004	Graziosi et al.27	Netherlands	RCT	2	Misoprostol, Cutterage	32.1	71.4	34
7	2020	Grewal et al.28	London	QCT	21	Expectant vs surgical	34	42 days	N/A
8	2019	Ibiyemi et al.29	Nigeria	RCT	7	Misoprostol vs surgery	28.38 (5.51)	N/A	N/A
9	2001	Karlsen et al.30	Norway	RCT	10	Expectant Management, Surgical Evacuation	30.8	59.5	1.1
10	2016	Lemmers et al.31	Netherlands	RCT	42	Cutterage, Expectant Management	31.8	N/A	16
11	2001	Ngai et al.32	China	RCT	15	Misoprostol vs expectant	31.5 (7.7)	43.5	14
12	2006	Niinimäki et al.33	Finland	RCT	30	Mifepristone+ misoprostol vs surgery	30.9 (6.9)	74.7	
13	2020	Nwafor et al.34	Nigeria	RCT	7	Misoprostol , MVA	N/A	58.8	1.6
14	2009	Prasad et al.35	India	RCT	8	Misoprostol vs surgery	N/A	48	N/A
15	2012	Shochet et al.36	Africa	Cohort	7 days	Surgical vs Medical	287	N/A	11
16	2013	Shokryet al.37	Egypt	Cohort	0.5	Misoprostol, Surgical Evacuation	27.1	58.8	11
17	2013	Shuaib et al.38	Yemen	RCT	7	Misopristole	28.9	N/A	43
18	2006	Trinder et al.39	United Kingdom	RCT	14	Misoprostol vs expectant vs Surgery	31.2 (5.9)	N/A	226
19	2002	Wieringa-de Waard, M et al.40	Amsterdam	QCT	42	Surgical Curettage, Expectant Management	32.8	54	14
20	2011	Wijesinghe et al.41	Sri Lanka	RCT	14	Expectant vs surgical	29.19 (5.67)	73.13 days	33 (46%)
21	2015	Zhang et al.42	US	RCT	84	Misoprostol, Surgical Evacuation	30.9	53.2	N/A

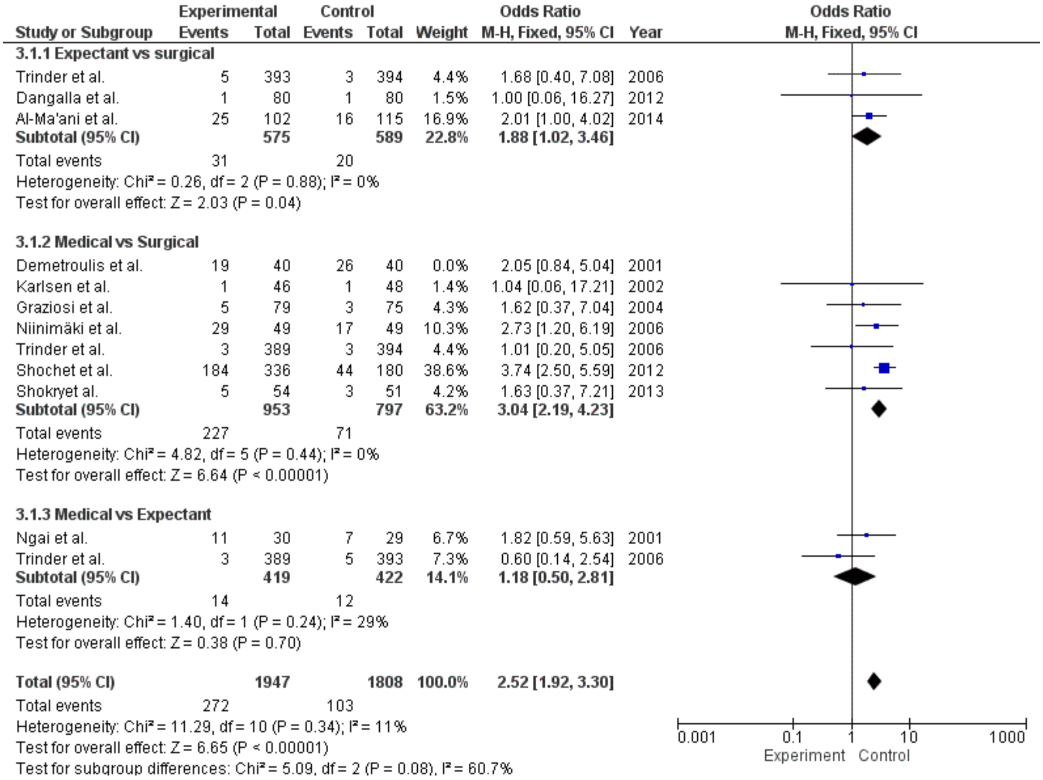
RCT: Randomized controlled trials; QCT: Quasi randomized controlled trials

Figure 2 a. Forest plot comparing success rate of surgical vs medical vs expectant



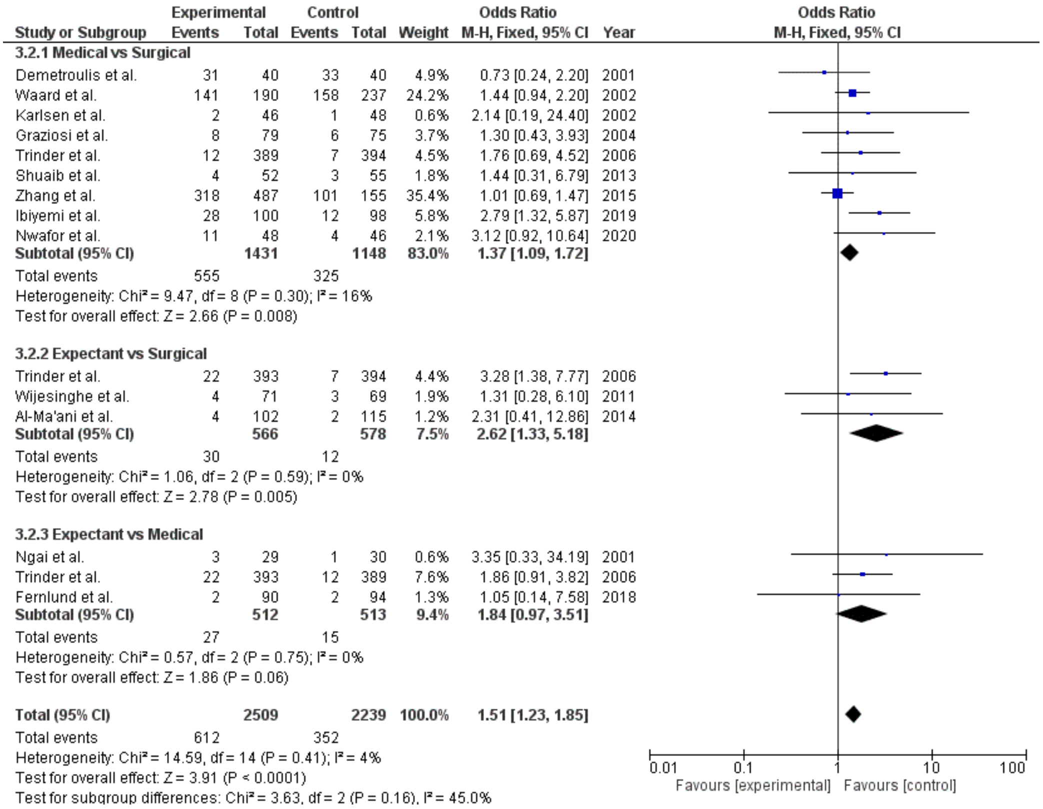
	Surgical		Control			Odds Ratio	
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**M-H, Fixed, 95% CI**	**Year**
1.1.1 Surgical vs expectant							
Trinder et al.	271	394	185	393	64.7%	2.48 [1.85, 3.31]	2006
Wijesinghe et al.	66	69	67	71	3.2%	1.31 [0.28, 6.10]	2011
Dangalla et al.	79	80	67	80	0.9%	15.33 [1.95, 120.25]	2012
Al-Ma'ani et al.	110	115	83	102	4.3%	5.04 [1.81, 14.04]	2014
Lemmers et al.	29	30	22	29	0.8%	9.23 [1.06, 80.60]	2016
Grewal et al.	7	9	18	26	2.3%	1.56 [0.26, 9.21]	2020
Subtotal (95% CI)		**697**		**701**	**76.3%**	**2.78 [2.13, 3.61]**	
Total events	562		442				
Heterogeneity: Chi² = 7.03, df = 5 (P = 0.22); I² = 29%							
Test for overall effect: Z = 7.59 (P < 0.00001)							
1.1.2 Surgical vs medical							
Demetroulis et al.	40	40	33	40	0.5%	18.13 [1.00, 329.27]	2001
Graziosi et al.	72	75	42	79	1.8%	21.14 [6.14, 72.81]	2004
Niinimäki et al.	47	49	45	49	2.1%	2.09 [0.36, 11.97]	2006
Bennett et al.	1148	1149	1262	1309	1.1%	42.75 [5.89, 310.38]	2010
Shochet et al.	360	362	393	420	2.2%	12.37 [2.92, 52.37]	2012
Shokryet al.	51	51	43	54	0.5%	27.23 [1.56, 475.50]	2013
Shuaib et al.	55	55	42	52	0.4%	27.42 [1.56, 481.27]	2013
Zhang et al.	97	98	83	100	0.9%	19.87 [2.59, 152.48]	2015
Ibiyemi et al.	97	98	83	100	0.9%	19.87 [2.59, 152.48]	2019
Nwafor et al.	44	46	39	48	1.9%	5.08 [1.03, 24.94]	2020
Subtotal (95% CI)		**2023**		**2251**	**12.4%**	**16.12 [9.11, 28.52]**	
Total events	2011		2065				
Heterogeneity: Chi² = 8.88, df = 9 (P = 0.45); I² = 0%							
Test for overall effect: Z = 9.54 (P < 0.00001)							
1.1.3 Medical vs expectant							
Ngai et al.	25	30	14	29	2.7%	5.36 [1.61, 17.88]	2001
Fernlund et al.	81	94	55	90	8.7%	3.97 [1.92, 8.17]	2018
Subtotal (95% CI)		**124**		**119**	**11.4%**	**4.29 [2.31, 7.97]**	
Total events	106		69				
Heterogeneity: Chi² = 0.18, df = 1 (P = 0.67); I² = 0%							
Test for overall effect: Z = 4.61 (P < 0.00001)							
Total (95% CI)		**2844**		**3071**	**100.0%**	**4.60 [3.71, 5.70]**	
Total events	2679		2576				
Heterogeneity: Chi² = 44.42, df = 17 (P = 0.0003); I² = 62%							
Test for overall effect: Z = 13.95 (P < 0.00001)							
Test for subgroup differences: Chi² = 30.19, df = 2 (P < 0.00001), I² = 93.4%							

Figure 2b. Forest plot comparing abdominal pain of surgical vs medical vs expectant



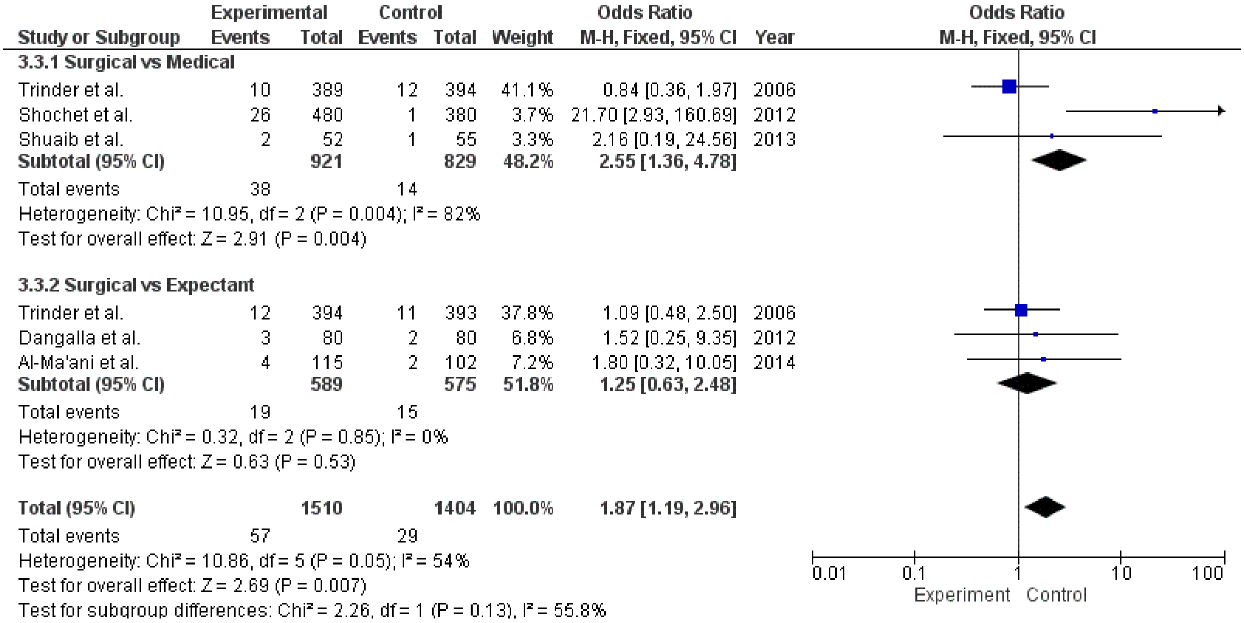
	Experimental		Control			Odds Ratio	
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**M-H, Fixed, 95% CI**	**Year**
3.1.1 Expectant vs surgical							
Trinder et al.	5	393	3	394	4.4%	1.68 [0.40, 7.08]	2006
Dangalla et al.	1	80	1	80	1.5%	1.00 [0.06, 16.27]	2012
Al-Ma'ani et al.	25	102	16	115	16.9%	2.01 [1.00, 4.02]	2014
Subtotal (95% CI)		**575**		**589**	**22.8%**	**1.88 [1.02, 3.46]**	
Total events	31		20				
Heterogeneity: Chi² = 0.26, df = 2 (P = 0.88); I² = 0%							
Test for overall effect: Z = 2.03 (P = 0.04)							
3.1.2 Medical vs Surgical							
Demetroulis et al.	19	40	26	40	0.0%	2.05 [0.84, 5.04]	2001
Karlsen et al.	1	46	1	48	1.4%	1.04 [0.06, 17.21]	2002
Graziosi et al.	5	79	3	75	4.3%	1.62 [0.37, 7.04]	2004
Niinimäki et al.	29	49	17	49	10.3%	2.73 [1.20, 6.19]	2006
Trinder et al.	3	389	3	394	4.4%	1.01 [0.20, 5.05]	2006
Shochet et al.	184	336	44	180	38.6%	3.74 [2.50, 5.59]	2012
Shokryet al.	5	54	3	51	4.2%	1.63 [0.37, 7.21]	2013
Subtotal (95% CI)		**953**		**797**	**63.2%**	**3.04 [2.19, 4.23]**	
Total events	227		71				
Heterogeneity: Chi² = 4.82, df = 5 (P = 0.44); I² = 0%							
Test for overall effect: Z = 6.64 (P < 0.00001)							
3.1.3 Medical vs Expectant							
Ngai et al.	11	30	7	29	6.7%	1.82 [0.59, 5.63]	2001
Trinder et al.	3	389	5	393	7.3%	0.60 [0.14, 2.54]	2006
Subtotal (95% CI)		**419**		**422**	**14.1%**	**1.18 [0.50, 2.81]**	
Total events	14		12				
Heterogeneity: Chi² = 1.40, df = 1 (P = 0.24); I² = 29%							
Test for overall effect: Z = 0.38 (P = 0.70)							
Total (95% CI)		**1947**		**1808**	**100.0%**	**2.52 [1.92, 3.30]**	
Total events	272		103				
Heterogeneity: Chi² = 11.29, df = 10 (P = 0.34); I² = 11%							
Test for overall effect: Z = 6.65 (P < 0.00001)							
Test for subgroup differences: Chi² = 5.09, df = 2 (P = 0.08), I² = 60.7%							

Figure 2c. Forest plot comparing Vaginal bleeding of surgical vs medical vs expectant



Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI	Year
3.2.1 Medical vs Surgical							
Demetroulis et al.	31	40	33	40	4.9%	0.73 [0.24, 2.20]	2001
Waard et al.	141	190	158	237	24.2%	1.44 [0.94, 2.20]	2002
Karlsen et al.	2	46	1	48	0.6%	2.14 [0.19, 24.40]	2002
Graziosi et al.	8	79	6	75	3.7%	1.30 [0.43, 3.93]	2004
Trinder et al.	12	389	7	394	4.5%	1.76 [0.69, 4.52]	2006
Shuaib et al.	4	52	3	55	1.8%	1.44 [0.31, 6.79]	2013
Zhang et al.	318	487	101	155	35.4%	1.01 [0.69, 1.47]	2015
Ibiyemi et al.	28	100	12	98	5.8%	2.79 [1.32, 5.87]	2019
Nwafor et al.	11	48	4	46	2.1%	3.12 [0.92, 10.64]	2020
Subtotal (95% CI)		**1431**		**1148**	**83.0%**	**1.37 [1.09, 1.72]**	
Total events	555		325				
Heterogeneity: Chi² = 9.47, df = 8 (P = 0.30); I² = 16%							
Test for overall effect: Z = 2.66 (P = 0.008)							
3.2.2 Expectant vs Surgical							
Trinder et al.	22	393	7	394	4.4%	3.28 [1.38, 7.77]	2006
Wijesinghe et al.	4	71	3	69	1.9%	1.31 [0.28, 6.10]	2011
Al-Ma'ani et al.	4	102	2	115	1.2%	2.31 [0.41, 12.86]	2014
Subtotal (95% CI)		**566**		**578**	**7.5%**	**2.62 [1.33, 5.18]**	
Total events	30		12				
Heterogeneity: Chi² = 1.06, df = 2 (P = 0.59); I² = 0%							
Test for overall effect: Z = 2.78 (P = 0.005)							
3.2.3 Expectant vs Medical							
Ngai et al.	3	29	1	30	0.6%	3.35 [0.33, 34.19]	2001
Trinder et al.	22	393	12	389	7.6%	1.86 [0.91, 3.82]	2006
Fernlund et al.	2	90	2	94	1.3%	1.05 [0.14, 7.58]	2018
Subtotal (95% CI)		**512**		**513**	**9.4%**	**1.84 [0.97, 3.51]**	
Total events	27		15				
Heterogeneity: Chi² = 0.57, df = 2 (P = 0.75); I² = 0%							
Test for overall effect: Z = 1.86 (P = 0.06)							
Total (95% CI)		**2509**		**2239**	**100.0%**	**1.51 [1.23, 1.85]**	
Total events	612		352				
Heterogeneity: Chi² = 14.59, df = 14 (P = 0.41); I² = 4%							
Test for overall effect: Z = 3.91 (P < 0.0001)							
Test for subgroup differences: Chi² = 3.63, df = 2 (P = 0.16), I² = 45.0%							

Figure 2d. Forest plot comparing Infection rate of surgical vs medical vs expectant



Study or Subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Weight	Odds Ratio M-H, Fixed, 95% CI	Year
3.3.1 Surgical vs Medical							
Trinder et al.	10	389	12	394	41.1%	0.84 [0.36, 1.97]	2006
Shochet et al.	26	480	1	380	3.7%	21.70 [2.93, 160.69]	2012
Shuaib et al.	2	52	1	55	3.3%	2.16 [0.19, 24.56]	2013
Subtotal (95% CI)		**921**		**829**	**48.2%**	**2.55 [1.36, 4.78]**	
Total events	38		14				
Heterogeneity: Chi² = 10.95, df = 2 (P = 0.004); I² = 82%							
Test for overall effect: Z = 2.91 (P = 0.004)							
3.3.2 Surgical vs Expectant							
Trinder et al.	12	394	11	393	37.8%	1.09 [0.48, 2.50]	2006
Dangalla et al.	3	80	2	80	6.8%	1.52 [0.25, 9.35]	2012
Al-Ma'ani et al.	4	115	2	102	7.2%	1.80 [0.32, 10.05]	2014
Subtotal (95% CI)		**589**		**575**	**51.8%**	**1.25 [0.63, 2.48]**	
Total events	19		15				
Heterogeneity: Chi² = 0.32, df = 2 (P = 0.85); I² = 0%							
Test for overall effect: Z = 0.63 (P = 0.53)							
Total (95% CI)		**1510**		**1404**	**100.0%**	**1.87 [1.19, 2.96]**	
Total events	57		29				
Heterogeneity: Chi² = 10.86, df = 5 (P = 0.05); I² = 54%							
Test for overall effect: Z = 2.69 (P = 0.007)							
Test for subgroup differences: Chi² = 2.26, df = 1 (P = 0.13), I² = 55.8%							

5 (P = 0.22); I^2 = 29%). Whereas medical intervention had an high success rate when compared with expectant management (N=243, OR: 4.29 [2.31, 7.97], Heterogeneity: Chi^2 = 0.18, df = 1 (P = 0.67); I^2 = 0%). The studies showed that risk of abdominal pain was higher in medical when compared to surgical (OR: 3.04 [2.19, 4.23]) and expectant management (OR: 1.18 [0.50, 2.81]) whereas the risk was higher in expectant compared to surgical (OR: 1.88 [1.02, 3.46]). The studies showed that risk of vaginal bleeding was higher in expectant group when compared with surgical (OR: 2.62 [1.33, 5.18]) or medical (OR: 1.84 [0.97, 3.51]), while there in increased risk in medical compared to surgical group (OR:1.37 [1.09, 1.72]). The rate of infection is higher in the surgical group when compared to medical (OR: 2.55 [1.36, 4.78]) and expectant group (OR: 1.25 [0.63, 2.48]).

3.4 Publication bias

The funnel plot was symmetrical, indicating absence of publication bias as shown in **Figure 3.** Which was confirmed using Egger's regression method[21] (Egger test, P=0.621).

DISCUSSION

Among the 21 selected studies, eleven studies compared medical intervention with surgical, three compared medical management with expectant management and eight studies compared surgical with expectant management for the management of spontaneous miscarriage in the first trimester. From the studies, it was observed that the success of complete abortion was higher in medical when compared to expectant whereas the medical treatment was inferior in comparison to surgical treatment. [26,39]

Mifepristone, an anti-progestin, works by blocking progesterone receptors, leading to softening and dilation of the cervix thus promoting the expulsion of pregnancy tissue. However, their effectiveness can vary depending on factors such as gestational age, dosage regimen, and individual patient response. [4,9] Surgical management use of suction or dilation and curettage (D&C) mechanically to scrape and remove tissue using surgical instruments provide direct and controlled removal of pregnancy tissue, ensuring a higher likelihood of complete abortion without the need for further intervention[4].

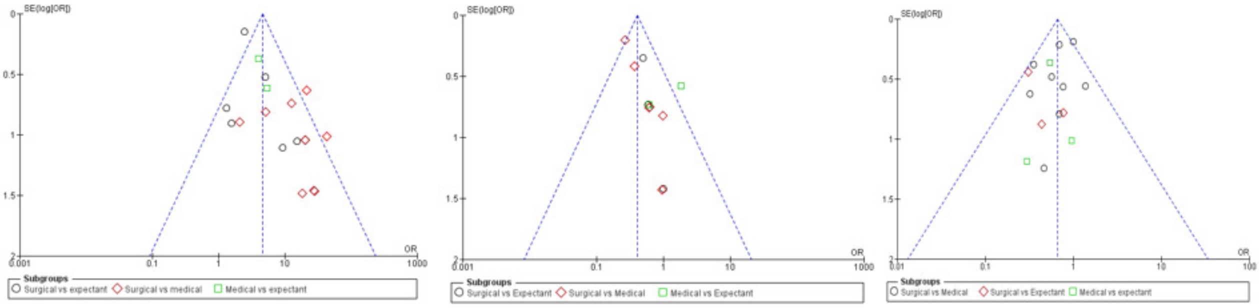
Though a higher success was observed in surgical trials, however the results of the trial a greater risk of infection following a surgical management with requirement for hospitalization when compared to medical or expectant management. Surgical methods involves invasive procedures that creates a channel for potential pathogens from the external environment or endogenous sources to enter the uterus, increasing the risk of infection.

Most common side effect observed in all three intervention was the risk of vaginal bleeding and abdominal pain among the patients before and after the management of miscarriage. The studies included collected history of vaginal bleeding and abdominal pain through self-report interviews or questionnaire. The pooled result of all the studies showed that the risk of vaginal bleeding was higher in the expectant group as this group needs to wait for the expulsion of the gestation tissue. The risk of abdominal pain was higher in the misoprostol group when compared to other intervention[43,44].

The risk of bias assessment of all the studies included in the systematic review was generally low to moderate. Blinding of participants and clinicians was not possible in most of the studies. There was no clarity regarding the selective reporting bias as the trial protocols were not assessed. Loss to follow-up and exclusions after randomization were low[45].

In present study we tried to minimize bias by assigning two independent reviewers to assess the eligibility for inclusion data extraction and assessed risk of bias independently. Data extraction was undertaken by one review author and checked by another. However, due to subjective assessments there might be some risk of bias.

Figure 3 a-c: Funnel plots of all individual studies in the meta-analysis



CONCLUSION

Although it would be critical to have more data, the current evidence suggests medical treatment is superior to expectant care in terms of success rate and less frequent side effects and can be an alternative to surgery management of first trimester miscarriage. Study has identified high risk of abdominal pain with the use of medical intervention, vaginal bleeding requiring blood transfusion in expectant management and higher infection rate in surgical group requiring hospitalization or antibiotic regimen. These side-effects should be explained to the women during treatment counselling. Further studies are required to compare the medical with expectant care. Future trials should consider women's views and quality of life measures alongside the clinical outcome.

DECLARATION

We confirm that the manuscript has been read and approved by all the listed authors. We further confirm that the order of authors listed in the manuscript has been approved by all.

Ethics approval and consent to participate: Ethical approval was not required for the present study as it is based on the secondary data/information.

Consent for publication: All the listed authors give their due consent for the publication

Availability of data and material: The present study is based on the secondary data sources which are available at mentioned databases in public domain. We have used the data from published articles for our research. Please refer Supplementary material S1.

Competing interests: There are no conflicts of interest declared by authors.

Funding: The authors acknowledge the Vritika Research Internship sponsored by the Science & Engineering Research Board (SERB), Department of Science and Technology (DST), Govt. of India, under Accelerate Vigyan Vritika scheme for carrying out the study.

Authors' contributions: Shivali Negi, Kavya Sharma, Anwesa Acharya, and Ananya Prabhu have contributed the data collection, analysis, and manuscript preparation. Ramesh Athe developed the study protocol, secured funds, supervised the study, and guided in manuscript preparation. Rinshu Dwivedi contributed to the development of study protocol and manuscript writing.

Acknowledgements: The authors express their sincere acknowledgement to the SERB Vritika and Indian Institute of Information Technology Dharwad (IIIT Dharwad) for their support and insightful inputs. The authors are grateful to the Department of Data Sciences and Intelligent Systems, IIIT Dharwad for their support and encouragement, which has helped in improving this study.

AI Statement: We confirm that the AI hasn't been used to prepare the manuscript and approved by all the listed authors.

REFERENCES

1. Li, Y.-T., Chen, F.-M., Chen, T.-H., et al., (2006). Concurrent Use of Mifepristone and Misoprostol for Early Medical Abortion. *Taiwanese Journal of Obstetrics and Gynecology*, *45*(4), 325–328. https://doi.org/10.1016/S1028-4559(09)60252-7
2. Li, Y.-T., Hsieh, J. C.-H., Hou, G.-Q., et al., (2011). Simultaneous use of mifepristone and misoprostol for early pregnancy termination. *Taiwanese Journal of Obstetrics and Gynecology*, *50*(1), 11–14. https://doi.org/10.1016/j.tjog.2010.09.002
3. Dimitriadis, E., Menkhorst, E., Saito, S., et al., (2020). Recurrent pregnancy loss. *Nature reviews. Disease primers*, *6*(1), 98. https://doi.org/10.1038/s41572-020-00228-z
4. Sotiriadis, A., Makrydimas, G., Papatheodorou, S., et al., (2005). Expectant, medical, or surgical management of first-trimester miscarriage: a meta-analysis. *Obstetrics & Gynecology*, *105*(5 Part 1), 1104-1113.
5. Shelley, J. M., Healy, D., & Grover, S. (2005). A randomised trial of surgical, medical and expectant management of first trimester spontaneous miscarriage. *The Australian and New Zealand Journal of Obstetrics and Gynaecology*, *45*(2), 122–127. https://doi.org/10.1111/j.1479-828X.2005.00357.x
6. Tunçalp, O., Gülmezoglu, A. M., & Souza, J. P. (2010). Surgical procedures for evacuating incomplete miscarriage. *The Cochrane Database of Systematic Reviews*, *2010*(9), CD001993. https://doi.org/10.1002/14651858.CD001993.pub2
7. Chung, T. K., Cheung, L. P., Leung, T. Y., et al., (1995). Misoprostol in the management of spontaneous abortion. *British Journal of Obstetrics and Gynaecology*, *102*(10), 832–835. https://doi.org/10.1111/j.1471-0528.1995.tb10852.x
8. Wai Ngai, S., Ming Chan, Y., Shan Tang, O., et al., (2001). Vaginal misoprostol as medical treatment for first trimester spontaneous miscarriage. In *Human Reproduction* (Vol. 16, Issue 7).
9. Kim, T. Y., Kim, S., & Schafer, A. L. (2020). Medical Management of the Postoperative Bariatric Surgery Patient. In K. R. Feingold (Eds.) et. al., *Endotext*. MDText.com, Inc.
10. Gülmezoglu, A. M., Villar, J., Ngoc, et al., (2001). WHO multicentre randomised trial of misoprostol in the management of the third stage of labour. *The Lancet*, *358*(9283), 689-695.
11. Brandner, P., Neis, K. J., Wagner, S., et al., (2001). Uterine and fetal findings at hysteroscopic evaluation of spontaneous abortions before D&C. *The Journal of the American Association of Gynecologic Laparoscopists*, *8*(4), 552–557. https://doi.org/10.1016/s1074-3804(05)60620-2
12. Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al.,

(2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ (Clinical research ed.)*, *372*, n71. https://doi.org/10.1136/bmj.n71

13. Moher, D., Shamseer, L., Clarke, M., et al., (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1
14. Athe, R., Rao, M. V., & Nair, K. M. (2014). Impact of iron-fortified foods on Hb concentration in children (<10 years): a systematic review and meta-analysis of randomized controlled trials. *Public health nutrition*, *17*(3), 579–586. https://doi.org/10.1017/S1368980013000062
15. Mendu, V. V. R., Nair, K. P. M., & Athe, R. (2019). Systematic review and meta-analysis approach on vitamin A fortified foods and its effect on retinol concentration in under 10 year children. *Clinical nutrition ESPEN*, *30*, 126–130. https://doi.org/10.1016/j.clnesp.2019.01.005
16. Joanna Briggs Institute (2011), Joanna Briggs Institute Critical Appraisal Checklist for Studies Reporting Prevalence Data, Joanna Briggs Institute, Adelaide.
17. DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2
18. Borenstein, M., Hedges, L., Higgin, J., & Rothstein, H. (2009). Introduction to meta-analysis. Hoboken, NJ: John Wily & Sons, Ltd.
19. Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.)*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557
20. Athe, R., Dwivedi, R., Pati, S., et al., (2020). Meta-analysis approach on iron fortification and its effect on pregnancy and its outcome through randomized, controlled trials. *Journal of family medicine and primary care*, *9*(2), 513–519. https://doi.org/10.4103/jfmpc.jfmpc_817_19
21. Egger, M., Davey Smith, G., Schneider, M., et al., (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical research ed.)*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629
22. Al-Ma'ani, W., Solomayer, E. F., & Hammadeh, M. (2014). Expectant versus surgical management of first-trimester miscarriage: A randomised controlled study. *Archives of Gynecology and Obstetrics*, *289*(5), 1011–1015. https://doi.org/10.1007/s00404-013-3088-1
23. Bennett, I. M., Baylson, M., Kalkstein, K et al., (2009). Early Abortion in Family Medicine: Clinical Outcomes. *The Annals of Family Medicine*, *7*(6), 527–533. https://doi.org/10.1370/afm.1051
24. Dangalla, D. P. R., & Goonewardene, M. R. (2012). Surgical treatment versus expectant care in the management of incomplete miscarriage: a randomised controlled trial. In *Ceylon Medical Journal* (Vol. 57).
25. Demetroulis, C. (2001). A prospective randomized control trial comparing medical and surgical treatment for early pregnancy failure. *Human Reproduction*, *16*(2), 365–369. https://doi.org/10.1093/humrep/16.2.365
26. Fernlund, A., Jokubkiene, L., Sladkevicius, P., et al., (2018). Misoprostol treatment vs expectant management in women with early non-viable pregnancy and vaginal bleeding: a pragmatic randomized controlled trial. *Ultrasound Obstet Gynecol*, *51*(1), 24–32. https://doi.org/10.1002/uog.18940
27. Graziosi, G. C. M. (2004). Misoprostol versus curettage in women with early pregnancy failure after initial expectant management: a randomized trial. *Human Reproduction*, *19*(8), 1894–1899. https://doi.org/10.1093/humrep/deh344
28. Grewal, K., Al-Memar, M., Fourie, H., et al., (2020). Natural history of pregnancy-related enhanced myometrial vascularity following miscarriage. *Ultrasound in Obstetrics and Gynecology*, *55*(5), 676–682. https://doi.org/10.1002/uog.21872
29. Ibiyemi, K. F., Ijaiya, M. A., & Adesina, K. T. (2019). Randomised Trial of Oral Misoprostol Versus Manual Vacuum Aspiration for the Treatment of Incomplete Abortion at a Nigerian Tertiary Hospital. *Sultan Qaboos Univ Med J*, *19*(1), e38–e43. https://doi.org/10.18295/squmj.2019.19.01.008
30. Karlsen, J. H., & Schiøtz, H. A. (2001). Curettage or not after spontaneous abortion?.*Tidsskrift for Den Norske Laegeforening : Tidsskrift for Praktisk Medicin, Ny Raekke*, *121*(24), 2812–2814.
31. Lemmers, M., Verschoor, M. A. C., Oude Rengerink, K., et al., (2016). MisoREST: surgical versus expectant management in women with an incomplete evacuation of the uterus after misoprostol treatment for miscarriage: a randomized controlled trial. *Human Reproduction*, *31*(11), 2421–2427. https://doi.org/10.1093/humrep/dew221
32. Ngai, S. W., Chan, Y. M., Tang, O. S., et al., (2001). Vaginal misoprostol as medical treatment for first trimester spontaneous miscarriage. *Hum. Reprod.*,*16*(7), 1493–1496. https://doi.org/10.1093/humrep/16.7.1493
33. Niinimäki, M., Jouppila, P., Martikainen, H., et al., (2006). A randomized study comparing efficacy and patient satisfaction in medical or surgical treatment of miscarriage. *Fertility and Sterility*, *86*(2), 367–372. https://doi.org/10.1016/j.fertnstert.2005.12.072
34. Nwafor, J., Agwu, U., Egbuji, C., et al., (2020). Misoprostol versus manual vacuum aspiration for treatment of first-trimester incomplete miscarriage in a low-resource setting: A randomized controlled trial. *Nigerian Journal of Clinical Practice*, *23*(5), 638–646. https://doi.org/10.4103/njcp.njcp_379_19
35. Prasad, S., Kumar, A., & Divya, A. (2009). Early termination of pregnancy by single-dose 800 μg misoprostol compared with surgical evacuation. *Fertility and Sterility*, *91*(1), 28–31. https://doi.org/10.1016/j.fertnstert.2007.11.028
36. Shochet, T., Diop, A., Gaye, A., et al., (2012). Sublingual misoprostol versus standard surgical care for treatment of incomplete abortion in five sub-Saharan African countries. *BMC Pregnancy and Childbirth*, *12*(1), 127. https://doi.org/10.1186/1471-2393-12-127

37. Shokry, M., Fathalla, M., Hussien, M., et al., (2014). Vaginal misoprostol versus vaginal surgical evacuation of first trimester incomplete abortion: Comparative study. *Middle East Fertility Society Journal*, *19*(2), 96–101. https://doi.org/10.1016/j.mefs.2013.05.007
38. Shuaib, A. A., & Alharazi, A. H. (2013). Medical versus surgical termination of the first trimester missed miscarriage. *Alexandria Journal of Medicine*, *49*(1), 13–16. https://doi.org/10.1016/j.ajme.2012.08.004
39. Trinder, J., Brocklehurst, P., Porter, R., et al., (2006). Management of miscarriage: Expectant, medical, or surgical? Results of randomised controlled trial (miscarriage treatment (MIST) trial). *Br. Med. J.*, *332*(7552), 1235–1238. https://doi.org/10.1136/bmj.38828.593125.55
40. Wieringa-de Waard, M. (2002). Management of miscarriage: a randomized controlled trial of expectant management versus surgical evacuation. *Human Reproduction*, *17*(9), 2445–2450. https://doi.org/10.1093/humrep/17.9.2445
41. Wijesinghe, P. S., Padumadasa, G. S., Palihawadana, T. S., et al., (2011). A trial of expectant management in incomplete miscarriage. *The Ceylon Medical Journal*, *56*(1), 10–13. https://doi.org/10.4038/cmj.v56i1.2888
42. Zhang, J., Gilles, J. M., Barnhart, K., et al., (2005). A Comparison of Medical Management with Misoprostol and Surgical Management for Early Pregnancy Failure. *New England Journal of Medicine*, *353*(8), 761–769. https://doi.org/10.1056/NEJMoa044064
43. Jewson, M., Purohit, P., & Lumsden, M. A. (2020). Progesterone and abnormal uterine bleeding/menstrual disorders. *Best practice & research. Clinical obstetrics & gynaecology*, *69*, 62–73. https://doi.org/10.1016/j.bpobgyn.2020.05.004
44. Ghana, S., Hakimi, S., Mirghafourvand, M., et al., (2017). Randomized controlled trial of abdominal binders for postoperative pain, distress, and blood loss after cesarean delivery. *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics*, *137*(3), 271–276. https://doi.org/10.1002/ijgo.12134
45. Tang, A. W., Alfirevic, Z., Turner, M. A., et al., (2013). A feasibility trial of screening women with idiopathic recurrent miscarriage for high uterine natural killer cell density and randomizing to prednisolone or placebo when pregnant. *Human reproduction (Oxford, England)*, *28*(7), 1743–1752. https://doi.org/10.1093/humrep/det117

Supplementary Table S 1

Population	(((((((((((((Miscarriage) OR (Recurrent spontaneous abortion)) OR (abortion)) OR (Recurrent pregnancy loss)) OR (Recurrent miscarriage)) OR (Spontaneous miscarriage)) OR (Spontaneous abortion)) OR (Pregnancy loss)) OR (Pregnancy)) OR (Pregnant)) OR (Gestation)) OR (1st trimester)) OR ("First-trimester"))
Intervention	((((((((((((((“Medical management”)) OR (Abortifacient Agents, Nonsteroidal)) OR (Abortifacient Agents)) OR (Misoprostol)) OR (Abortifacient agents, steroidal)) OR (Mifepristone)) OR (chorionic gonadotropin)) OR (oxytocics)) OR (“Medical”)) OR (prostaglandin analogue)) OR (mifepristone)) OR (antiprogesterone))) AND (((((((("Expectant management")) OR (Monitoring)) OR (Active monitoring)) OR (Waiting, watchful)) OR (Management, expectant)) OR (Active surveillance)) OR (Surveillance, active)) OR (Follow ups)))) AND (((((((((((("Vacuum Curettage") OR ("Vacuum Extraction, Obstetrical")) OR ("Operative hysteroscopy")) OR ("Ambulatory Surgical Procedures")) OR ("Dilatation and Curettage")) OR ("Electric vacuum aspiration")) OR ("Manual vacuum aspiration")) OR ("Suction aspiration")) OR (Surgery*)) OR ("Surgical management")) OR ("Surgical treatment")) OR (Curettage*)))
Outcome	(((((((((((((((((((((((((((((Haemorrhage) OR (Blood loss)) OR (Excessive blood loss)) OR (Excessive bleeding)) OR (Bleeding)) OR (Uterine haemorrhage)) OR (Uterine hemorrhage)) OR (hemorrhage)) OR (Blood transfusion)) OR (Febrile morbidity)) OR (Post-operative Febrile morbidity)) OR (Post-operative fever)) OR (Fever)) OR (High fever)) OR (Repeated uterine evacuation)) OR (Uterine evacuation)) OR (Repeated surgical evacuation)) OR (Second uterine evacuation)) OR (Incomplete uterine evacuation)) OR (Reinfection)) OR (Gynaecological infection)) OR (Rehabilitation)) OR (Rehospitalisation)) OR (Rehospitalization)) OR (Post operative pain)) OR (Abdominal pain)) OR (Post operative abdominal pain)) OR (Antibiotic medication)) OR (Antibiotic therapy)) OR (Antibiotic drugs)) OR (Chemotherapy)

Table S2. Quality Assessment by using JBI checklists.

Authors	**Study design**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Q13**
Ngai et al.	RCT	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Niinimäki et al.	RCT	Yes	Yes	No	N/A	Unclear	unclear	Yes	No	Yes	Yes	Yes	Yes	Yes
Trinder et al.	RCT	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Prasad et al.	RCT	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Unclear	Yes	Yes	Yes
Shuaib	RCT	Yes	Yes	Unclear	No	No	No	Yes	Yes	Yes	No	Unclear	Yes	Yes
Fernlund et al.	RCT	Yes	Yes	No	No	No	No	Yes	No	Yes	Yes	No	No	No
Ibiyemi et al.	RCT	Yes	Yes	No	No	No	Unclear	Yes	Unclear	Yes	Unclear	No	Yes	Yes
Wijesinghe et al.	RCT	Yes	Yes	Yes	No	No	Unclear	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dangalla et al.	RCT	Yes	Yes	Yes	No	No	Unclear	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Al-Ma'ani et al.	RCT	Yes	Yes	Yes	No	No	Unclear	Yes	No	Yes	Yes	Yes	No	Yes
Nwafor et al.	RCT	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Zhang et al.	RCT	Yes	Yes	Yes	Unclear	Unclear	Unclear	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lemmers et al.	RCT	Yes	Yes	Yes	No	No	No	No	Unclear	Yes	Yes	Yes	Yes	Yes
Graziosi et al.	RCT	Yes	Yes	Yes	Unclear	Unclear	Unclear	Yes	Unclear	Yes	Yes	Yes	Yes	Yes
Demetroulis et al.	RCT	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Karlsen et al.	RCT	Yes	Yes	Yes	Yes	Unclear	Unclear	Yes	Yes	Yes	Yes	Yes	Yes	Yes
		Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**		
Bennett et al.	Cohort	Yes	Unclear	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Yes		
Shokryet al.	Cohort	Yes	Yes	Unclear	Yes	No	No	Yes	Yes	Unclear	No	Yes		
Shochet et al.	Cohort	Yes	Yes	Yes	Unclear	Unclear	Unclear	No	Yes	Yes	Yes	Yes		
		Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**				
Grewal et al.	QCT	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes				
Waard et al.	QCT	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes				

RCT: Randomized Controlled Trials, QCT: Quasi Randomized Controlled Trials

ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - 2024, Volume 19, Issue 1

Impact of COVID-19 Lockdown on Air Pollutant Emissions in Port Regions. Scoping Review

Sandro Roberto Mastellari Francisco(1), Giullia Carvalho Mangas Lopes(2), Gustavo Duarte Mendes(3), Nicole Amado Riso(4), Victória Iglesias Breda(4), Marcela Gonçalves Leal(3), Elaine Marcílio Santos(3), Ana Luiza Cabrera Martimbianco(3,5)

(1) Ms, Postgraduate Program in Health and Environment, Universidade Metropolitana de Santos, Santos, Brazil.
(2) Master's student, Postgraduate Program of Health and Environment, Universidade Metropolitana de Santos, Santos, Brazil.
(3) PhD, Professor, Postgraduate Program of Health and Environment, Universidade Metropolitana de Santos, Santos (Unimes), Brazil.
(4) Medical student, Universidade Metropolitana de Santos, Santos, SP, Brazil.
(5) Centre of Health Technology Assessment, Hospital Sírio-Libanês, São Paulo, Brazil.

CORRESPONDING AUTHOR: Ana Luiza Cabrera Martimbianco, Postgraduate Program of Health and Environment, Universidade Metropolitana de Santos, Santos, SP, Brazil. Avenida Conselheiro Nébias 536, Santos - SP, Brazil. Zip code 11045-002. Phone: +55 (13) 3228-3400. E-mail: analuizacabrera@hotmail.com

SUMMARY

Background: This scoping review, a comprehensive effort to map and synthesize evidence, sheds light on the impacts of the COVID-19 lockdown on air pollutant emissions in port regions.
Methods: It was conducted based on the Joanna Briggs Institute Manual and the PRISMA-ScR recommendations. An extensive literature search was undertaken to identify any scientific study or report comparing greenhouse gas emissions before and after the COVID-19 pandemic in maritime port regions.
Results: Nine observational studies conducted in ports of five countries were identified, 75% using the Automatic Identification System (AIS) as a measurement system for pollutant emissions. When comparing the same period before and after the pandemic lockdown, the results of seven studies identified a reduction of up to 63% in the emission of CO_2, NO_2, CO, HC, NO_x, SO_x, HC, $PM_{2.5}$, and PM_{10}, and an increase of 37% in O_3. Additionally, two studies reported increased pollutant emissions, explained by ship congestion in ports.
Conclusion: These findings indicate an important reduction in pollutant and particulate matter emissions during the port activity restrictions imposed by the COVID-19 pandemic worldwide compared to the same period in 2019. This reduction was mainly attributed to the reduced activity of vessels and vehicle circulation. These findings can provide valid scientific evidence to support the air pollution control policies in coastal cities and assist in ensuring sustainable practices, environmental regulations, and monitoring for mitigating air pollution in port regions.

Keywords: Port activities; Atmospheric pollutants; COVID-19; Scoping review.

INTRODUCTION

Maritime transport, a cornerstone of global trade and development for almost 70 years, plays an indispensable role in the world economy. It facilitates the movement of approximately 10 billion tons of cargo, passengers, and crew annually, as well as more than 80% of the world's transported goods. However, maritime transport poses critical environmental challenges, especially with air pollutants and greenhouse gas emissions [1,2]. As stated by statistics from the United Nations Conference on Trade and Development (UNCTAD), there were approximately 105,500 ships in the global maritime sector at the beginning of 2023 compared to 92,295 in 2019. The majority rely on petroleum-derived fuels such as diesel and heavy fuel oil, which release large atmospheric pollutants [3,4]. According to the International Maritime Organization (IMO), shipping emitted approximately 1.056 billion tonnes of carbon

DOI: 10.54103/2282-0930/24059
Accepted: 31th July 2024

dioxide (CO_2) in 2018, representing approximately 2.89% of global greenhouse gas emissions. It is estimated that emissions from maritime transport will double by 2050 [5].

Associated with maritime activity and due to the great demand for the transport of goods, the regions surrounding seaports also concentrate high emissions of atmospheric pollutants. The concentration of ships, cargo handling equipment, local traffic with excess trucks, and other industrial activities in port areas worsens air pollution. It represents a higher risk of environmental damage and directly harms human health, including the higher prevalence of respiratory and cardiovascular diseases and reduced quality of life for local communities [6,7,8].

Since the start of the COVID-19 (Coronavirus disease 2019) pandemic in 2020, several health restrictions have been imposed worldwide to mitigate the spread of the virus. This scenario has led to an unprecedented global hiatus in several sectors, including the paralysis of economic activities and reduced human mobility [6,9]. One notable area of change has been the alteration of air pollutant emissions in port regions. During the lockdown, maritime transport activities were significantly reduced, and many industries located in or near port regions were closed or started to operate at reduced capacity, thus restricting the movement of cargo vehicles and human activities. Despite the emergency, such changes have provided a rare opportunity to identify and better understand the drivers of global environmental disruption, including the impact on air pollution in port regions [6,9].

The growing global interest related to changes in air quality has led to the significant production of scientific studies on port regions around the world to compare the emission of air pollutants before and after the COVID-19 lockdown. Therefore, it is required to investigate the results derived from this evidence to understand the mechanism of pollution sources and their relationship with the health of the population and the environment to support the implementation of preventive and sustainable measures to promote cleaner technologies and improve environmental regulations in port regions. Therefore, this scoping review aimed to map and synthesise results from studies that evaluated the impacts of restrictions imposed by the COVID-19 pandemic on maritime port activities regarding atmospheric emissions.

METHODS

This scoping review was planned and conducted based on the recommendations of the Joanna Briggs Institute Manual for scoping reviews [10]. The review report adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses - extension for scoping reviews (PRISMA-ScR) guidelines [11]. The review protocol was registered in the Open Science Framework (OSF) platform (available at https://osf.io/cknhv/) [12].

The research question was structured using the PCC acronym as follows:

- P (population): studies evaluating the environmental and/or health impacts of the COVID-19 pandemic on the port population, including port workers (permanent or temporary) and the population residing in port regions and surroundings.
- C (concept): analyses of the impact of COVID-19 pandemic-related restrictions on port activities and their environmental effects on atmospheric pollutant emissions. Approaches for measuring atmospheric pollutant emissions before and after the COVID-19 pandemic.
- C (context): seaport regions and surrounding areas, any country.

Eligibility criteria

We planned to include any primary (analytical or descriptive observational studies) or secondary study design, such as narrative or systematic reviews, that assessed the atmospheric pollutant emissions in seaport regions and surrounding areas before and after the COVID-19 pandemic. Full publications or abstracts presented at conferences and events were considered for inclusion.

Sources of information

A comprehensive search was conducted through structured search strategies for the following databases on November 12, 2023: Medical Literature Analysis and Retrieval System Online (MEDLINE, via PubMed), EMBASE (via Elsevier), Cochrane Library (via Wiley), Biblioteca Virtual em Saúde (BVS), Epistemonikos, Health Systems Evidence, SCOPUS, and WHO-COVID.

We also searched the following grey literature databases and preprint repositories: Data Archiving and Networked Services (DANS) and Open Science Preprints.

Additional unstructured searches were carried out on the following websites:

- International Maritime Organization (https://www.imo.org/en/OurWork/IIIS/Pages/Port%20State%20Control.aspx)
- World Ports COVID-19 Information Portal (https://sustainableworldports.org/world-ports-covid19-information-portal/)
- Port Economics (https://www.porteconomics.eu/category/thema/ports-covid-19/)
- Port Economics, Management and Policy (https://porteconomicsmanagement.org/pemp/contents/part9/ports-and-pandemic/)
- European Maritime Safety Agency (https://www.emsa.europa.eu/)
- MEDPorts Association (https://medports.org/)

- American Association of Port Authorities (www.aapa-ports.org)
- Centers for Disease Control and Prevention (https://www.cdc.gov/)
- McMaster Daily News COVID-19 (https://covid19.mcmaster.ca/)
- Oxford COVID-19 Evidence Service (https://www.cebm.net/oxford-covid-19-evidence-service/)
- World Health Organization (WHO) Coronavirus disease (COVID-19) pandemic (https://www.who.int/emergencies/diseases/novel-coronavirus-2019)

Manual searches were conducted in relevant reference study lists, and experts in the field were contacted. No language filter was applied, and there were no restrictions on date or language. The search strategies for each database and information source are detailed in Supplementary Material 1.

Study selection process

The study selection process used the Rayyan platform in two phases [13]. The first phase involved reading the titles and abstracts of the references found by the search strategies, and the second phase involved reading the full text of "potentially eligible" studies to confirm eligibility. Justifications for excluding studies at this stage were reported. Two independent reviewers conducted both phases, and a third reviewer resolved discrepancies in decisions to include or exclude studies.

Data extraction

Two reviewers independently extracted data from studies identified and included in this review, and discrepancies in information were resolved through consensus. The following data were collected for each study: publication year, study design, publication status (full article or abstract), study funding sources, atmospheric pollutants, and their monitoring systems, such as the Automatic Identification System (AIS), which is required on all vessels with gross tonnage, i.e., capacity volume exceeding 300 tons; and pollutant emission estimation model, capable of adjusting fuel consumption based on volume, power, speed, and the dynamic behaviour and operating mode of the vessel during a specified period. The AIS provides real-time data on the evolution of port calls (ship stops within the port) and maritime traffic (movement of vessels within and beyond the port's 30 nautical miles range). Authors of included studies could be contacted if additional information was needed.

Data synthesis and presentation

The qualitative synthesis of included studies was presented using a narrative approach and in graphs and tables with descriptive statistics (percentage and mean/standard deviation) related to atmospheric pollutant emission concentrations in different seaport scenarios.

RESULTS

Search Results

Search strategies retrieved 3641 references. After removing 203 duplicates, 3438 references were selected for the screening process through title and abstract analysis. After eliminating 3428 references that did not meet the inclusion criteria, 10 were identified as potentially eligible studies. The full texts were analysed, and one study was excluded as it solely assessed the economic impact of the COVID-19 pandemic on maritime transportation [14]. In the end, nine studies were included in the review [15-23]. The PRISMA flowchart of the study selection process is represented in Figure 1.

Characteristics of Included Studies

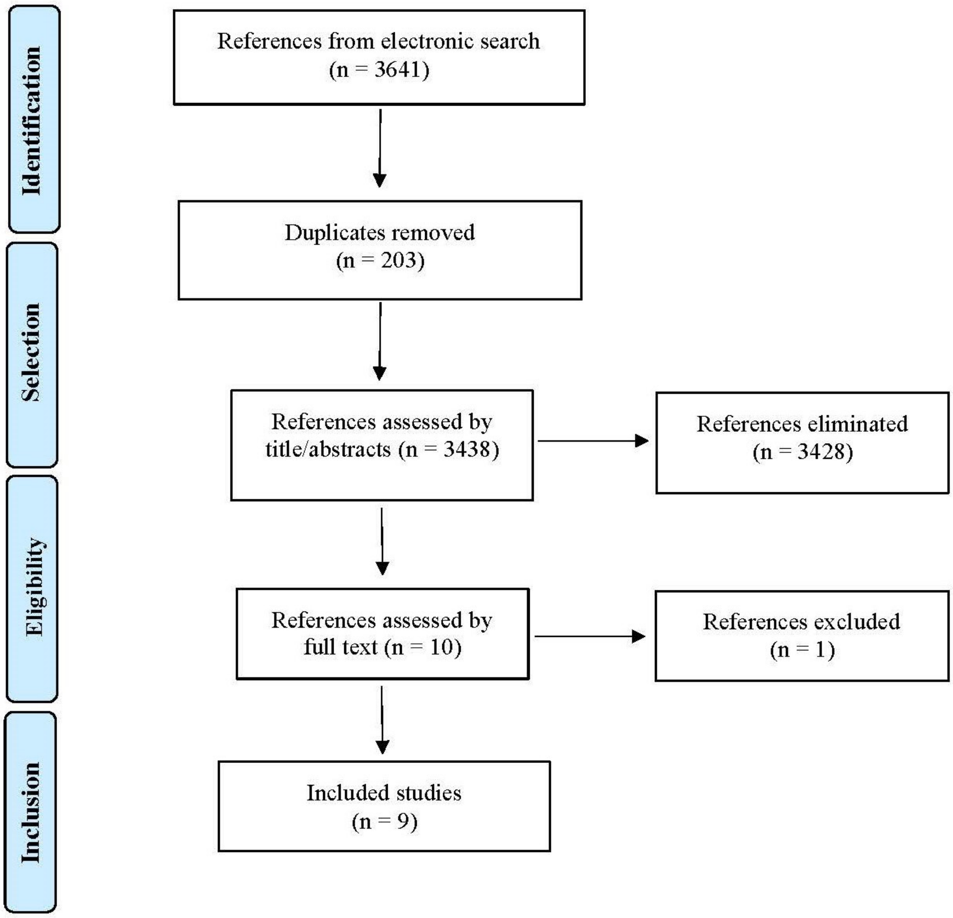
The main characteristics of the included studies are detailed in Table 1. All studies had a comparative observational design and were published as full articles. The studies were published between 2021 and 2023 and assessed, among other outcomes, the impact of sanitary restrictions imposed by the COVID-19 pandemic on the emission of atmospheric pollutants in ports from various countries, including 33,3% from China [15,17,23], 22,2% from Spain [16,21], and 44,5% from each of the following countries: Singapore [19], Italy [20], Brazil [22], and USA [18].

The monitoring systems varied among the studies; however, 75% utilised the AIS. Based on the data obtained, the studies adopted some models to estimate pollutant emissions from all vessels, such as the STEAM algorithm (Ship Traffic Emission Assessment Model) [16,21] or the MEET (Methodologies for Estimating Air Pollutant Emissions from Transport) and TRENDS (Transport and Environment Database System) models [19].

Furthermore, the studies reported monitoring and controlling climate variables and meteorological conditions, including temperature, humidity, wind speed and direction, precipitation, and thermal inversion episodes. The SENEM model (Ship's Energy Efficiency Model), used by Durán-Grados et al. (2020) [16], predicts speed loss due to additional resistance in abnormal weather conditions (irregular waves and wind).

Most of the included studies evaluated emission factors through calculations for the following pollutants: carbon dioxide (CO_2) and sulphur dioxide (SO_2), related to vessel fuel; nitrogen oxides (NO_x), related to the engine's type (main and auxiliary); and particulate matter (PM 2.5 and 10), among others, such as carbon monoxide (CO) and ozone (O_3).

Figure 1: PRISMA Flowchart of the study selection process


Identification
Selection
Eligibility
Inclusion
References from electronic search (n = 3641)
Duplicates removed (n = 203)
References assessed by title/abstracts (n = 3438)
References eliminated (n = 3428)
References assessed by full text (n = 10)
References excluded (n = 1)
Included studies (n = 9)

Results of included studies

Table 2 presents the findings from the nine included studies, which compared the concentration of air pollutant emissions in seaport regions before and after the COVID-19 pandemic lockdown period.

Seven studies [15,16,18,19-21,23] provided numerical data (in percentage) for the comparative analysis of pollutant gas emissions between the pre-pandemic (2019) and post-pandemic (2020/2021) periods. Regarding NO_2, only one study [16] reported the percentage variation between the periods, showing a 63% reduction when comparing them. The same study showed an increase in O_3 emissions by 38% in 2020 compared to 2019. Ultimately, the studies by Gu & Liu (2023) [17] and He et al. (2023) [18] also identified an increase in concentrations of MP2.5, MP10, SO_2 e O_3 explained by the port congestion scenario, where the increase in vessel berthing time occurred due to the quarantine period, sanitary restrictions, and port logistics blockade

Table 1: Main characteristics of the included studies

Study, year	Location	Monitoring system / Emission estimation model	Measured air pollutant concentrations	Sampling period	Funding sources
Chen, 2021	Port of Ningbo, China	Electrochemical sensor package ('home-made')	NO_2, O_3 e CO	January to March 2020 *versus* February to March 2019	National Key Research and Development of China and Natural Sciences Foundation of Zhejiang Province, China
Durán-Grados, 2020	Port of Algeciras and Strait of Gibraltar, Spain	AIS system / STEAM and SENEM	CO_2, CO, NO_x, SO_x HC, NO_2, NMVOC, PM	90 days during the COVID-19 lockdown *versus* 'no pandemic period'	Ministry of Health, Andalusia, Spain, and other institutions
Gu, 2023	14 ports from China	AIS system / Panel Data Regression Model	$PM_{2.5}$, PM_{10}, CO, NO_2, SO_2	January 2020 to July 2021 *versus* 'no pandemic period'	National Social Science Foundation of China
He, 2023	Port of Long Beach, USA	AIS system / Bottom-up method	CO_2, CO, NO_x, SO_x, PM	2019 to 2021	Nanyang Technological University, Singapore.
Ju, 2021	Port of Singapore, Singapore	AIS system / MEET and TRENDS	CO_2	2020 *versus* 'no pandemic period'	No financial support reported
Mocerino, 2021	Port of Naples, Italy	AIS system/ MEET	NO_X, SO_X, PM	2020 *versus* 2019	No financial support reported
Mujal-Colilles, 2022	Port of Barcelona, Spain	AIS system / STEAM	CO_2, SO_2, PM, NO_x	March to June 2020 *versus* 2019	IAMU and ACCI´O, Spanish MINECO program; European Research Council
Sarra, 2022	Port of Santos, Brazil	CETESB Monitoring / CETESB-QUALAR	PM, NO_x, SO_X	2020 *versus* 2019	No financial support reported
Shi, 2021	Port of Shanghai, China	AIS system / Botton-up method	CO_2, CO, NO_x, SO_x HC, NO_2, PM	2020 *versus* 2019	National Natural Science Foundation of China

AIS: Automatic Identification System; CETESB: Environmental Company of the State of São Paulo; MEET: Methodologies for Estimating air Pollutant Emissions from Transport; NMVOC: non-methane volatile organic compounds; SENEM: Ship's Energy Efficiency Model; STEAM: Ship Traffic Emission Assessment Model; TRENDS: TRansport and Environment Database System. CO_2: carbon dioxide, SO_2: sulfur dioxide, NO_x: nitrogen oxides, PM: particulate matter, CO: carbon monoxide, O_3: ozone, HC: hydrocarbons.

Table 2: Results of included studies regarding pollutant emissions in Port regions

Study, year	Location	Measurement period	Results
Chen, 2021	Port of Ningbo, China	2019 *versus* 2020	• Mean NO_2 = 19.5 *versus* 7.2 ppb (- 63%) • Mean O_3 = 27.5 *versus* 7.5 ppb (+ 38%) • Mean CO = 696.6 *versus* 648.5 ppb (- 7%)
Durán-Grados, 2020	Port of Algeciras and Strait of Gibraltar, Spain	90 days during the COVID-19 lockdown *versus* 'no pandemic period'	• 10% to 12% reduction in all the pollutants emissions, compared to no pandemic period
Gu, 2023	14 ports, China	January 2020 to July 2021 *versus* 'no pandemic period'	• Increase concentrations of $PM_{2.5}$, PM_{10}, SO_2 e O_3 (percentage data not reported), compared to no pandemic period
He, 2023	Port of Long Beach, USA	2019 *versus* 2021	• Overall pollutant emissions increase from 68% to 85% due to port congestion in 2021 • Reduced CO_2 emission: 52,527 *versus* 46,687 million tonnes
Ju, 2021	Port of Singapore, Singapore	2020 *versus* 'no pandemic period'	• 11% reduction in CO_2, compared to no pandemic period
Mocerino, 2021	Port of Naples, Italy	2019 *versus* 2020	• NO_x 332 t *versus* 62 t (- 18%) • SO_x 12,6 t *versus* 2,4 t (- 19%) • PM 23,5 t *versus* 4,4 t (- 18%)
Mujal-Colilles, 2022	Port of Barcelona, Spain	2019 *versus* 2020	• NO_x - 1,3% • CO_2- 1,8% • No significant reduction in SO and PM
Sarra, 2022	Port of Santos, Brazil	2019 *versus* 2020	• Operational activities (cargo handling) increased by 9.4% - bulk grain loading operation • Pollutant emissions in 2020 compared to 2019 (numerical data not reported): Increase in PM_{10} Reduction in $PM_{2.5}$ Reduction in SO_2 Reduction in NO_x
Shi, 2021	Porto de Shanghai, China	2019 *versus* 2020	• Ship count: 7.770 *versus* 4.085 (- 52,5%) • Average berthing time: 1,16h *versus* 4,64h (larger ships > time) • Reduction in pollutant emissions in 2020 compared to 2019: 14,7% CO_2 16,6% CO 10,5% HC 13,3% NO_x 16,3% $PM_{2.5}$ 16,2% PM_{10} 15,7% S_O2

CO_2: carbon dioxide; CO: carbon monoxide; HC: hydrocarbons; NO_x: nitrogen oxides; $PM_{2.5}$: particulate matter with a diameter less than 2.5 micrometers; PM_{10}: particulate matter with a diameter less than 10 micrometers; SO_2: sulfur dioxide; SO_x: sulfur oxides; AA: auxiliary engine; ME: main engine; μg/m3: micrograms per cubic meter; t: tons.

DISCUSSION

This scoping review was developed to identify and synthesise available evidence on the impacts of restrictions imposed by the COVID-19 pandemic on port activities concerning atmospheric pollutant emissions. Nine descriptive observational studies published between 2020 and 2023 were identified, analysing the emission of pollutant gases in ports of five countries and comparing periods before and after the COVID-19 pandemic. Seven studies showed a reduction of around 7% to 63% in the emission of CO_2, NO_2, CO, HC, NO_x, SO_x HC, MP2.5, and MP10, and an increase of 38% in O_3, which is closely related to the decrease in NO_2 concentration in a scenario of volatile organic compounds. This variability in emissions is probably attributable to factors such as the varying sizes of the ports, their potential for maritime operations, and the number of vessels they accommodate. Additionally, two studies [17,18] observed increased greenhouse gas emissions explained by port congestion during COVID-19 restrictions. Conversely, one study [23] identified a reduction of over 50% in the number of ships in the port of Shanghai, likely due to large-scale suspension measures taken by cargo transport companies to mitigate pandemic-induced losses. Notably, the restrictive measures adopted to contain the pandemic led to decreased vehicle circulation, such as heavy trucks, in the port area and surroundings, likely enhancing the observed pollutant emission reduction in the studies.

The measurement systems for atmospheric pollutant emissions from different vessels consider energy demand and, therefore, the ship's dynamic behaviour and operation mode. Thus, emission models related to maritime transport within a spatial region typically consider separate emission modes for each vessel, i.e., cruising, manoeuvring, hoteling, and berthing [21].

Other factors affecting the estimates of air pollution from ships are related to engine and fuel type [18]. The total fuel consumption and emissions that ships generate depend on the vessel type and the emission mode under which the engine operates. When the ship cruises, the main auxiliary engine works while the boiler is closed. The main engine, auxiliary engine, and boiler will operate simultaneously when the ship is in manoeuvring and anchoring conditions. However, when the ship is in berthing operation, the main engine is turned off while the auxiliary and the boiler are still operating [23]. In some countries, the berthing time was longer due to strict quarantine measures, resulting in more CO_2 and SO_2 produced by the boilers of merchant ships, even though the greater volume of pollutants comes mainly from the main engines [17]. Hence, the pandemic may have altered the emission distributions of vessels due to changes in their activities.

Assigning the appropriate emission mode is crucial for a more accurate emission estimate. Although the AIS system provides navigation status information, it is a fixed variable on the vessel. The ship's crew manually changes the AIS status and is therefore vulnerable to human errors and delays. Thus, the model used to estimate vessel emissions should consider the vessel's speed and location (within or outside port facilities) and the navigation status provided by AIS data. Another critical point is the relationship between passenger ships (cruise tourism) and pollutant emissions. These vessels operate constantly with a large cargo volume and high speeds, accounting for a significant portion of air pollution despite representing a smaller proportion of maritime traffic [21].

There is a strong correlation between the vessel's operation mode and its overall contribution to pollution. This aligns with the fact that more environmentally friendly navigation can only be achieved by reducing the average speed of vessels. The emission factor, the most critical parameter for estimating pollutant emissions, can be affected by various external factors, including engine types (main engine, auxiliary engine, boiler), fuel types (residual oil, marine distillate oil, marine gas oil), and engine status (low-speed diesel, medium-speed diesel) [23].

The increase in particulate matter emissions is related to the movement and loading of solid bulk cargo, such as soybeans and corn. One study [22] conducted at the port of Santos showed an increase in the proportion of solid bulk from 2019 to 2020, from 49% to 51.6%, with an increase in PM_{10} averages during the same period accompanied by a reduction in $PM_{2.5}$ concentrations. According to the authors, the higher movement of solid bulk tends to decrease the $PM_{2.5}/PM_{10}$ ratio due to the increased emission of PM_{10}.

Another critical factor in calculating ship emissions is considering weather and meteorological conditions that can influence the dispersion of gases and particulate matter, including the direction and speed of wind, direction and height of waves, thermal inversions, and precipitation. The study of wind behaviour is crucial as it affects pollutant dispersion conditions. Rainy weather, for example, reduces concentrations of atmospheric pollutants and increases air humidity, directly affecting sensor accuracy [19,20]. In this review, all included studies considered these measurements in their estimates of pollutant emissions.

The strengths of this scoping review involve the broad and sensitive search across various general databases and sources of information related to maritime port activities. Other aspects that provide methodological robustness involve the selection and data extraction performed in duplicate and adopting methods recommended by the Joanna Briggs Institute Manual for scoping reviews [11]. Limitations are related to the data obtained from the included studies, mainly due to possible imprecision of some monitoring systems and algorithms used for pollutant emission estimation and differences between port policies and pandemic restrictive measures among the analysed countries. Additionally, although there appears to be a relationship between pollutant emissions, fuel consumption, and the number of vessels in the region, several other factors play an essential role in the final values of air quality, including vehicle traffic, which was reduced during the pandemic. No similar scoping reviews were found.

This scoping review provides some critical insights into understanding the effects of port activities on health and the environment, developing practical implications, and identifying preventive and mitigating strategies for atmospheric pollution in coastal cities. Fundamental measures need to be taken to improve the efficiency of port logistics and air quality in port cities. Port infrastructure and handling equipment should be modernised to enhance operational efficiency [24]. Furthermore, port authorities and stakeholders should strengthen cooperation between sectors to modify sustainability environmental policies and legislation, promote more environmentally friendly navigation, rigorously monitor companies by environmental agencies, and quantify pollutants and particulate matter through reliable real-time data systems. Implementing Green Port concepts [25] is also essential and has implications for future research. Prospective and high-quality methodological studies are needed to analyse and monitor the impact of atmospheric pollution on the health of port workers and the resident population to provide evidence to support strategies to mitigate air pollution in seaport areas.

CONCLUSION

The findings of this scoping review showed a reduction in pollutant and particulate matter emissions during the period of port activity restrictions imposed by the COVID-19 pandemic worldwide, compared to the same period in 2019. This reduction was mainly attributed to the reduced activity of vessels and vehicle circulation. These findings can provide valid scientific evidence to support the air pollution control policies in coastal cities and assist in ensuring sustainable practices, environmental regulations, and monitoring for mitigating air pollution in port regions.

Competing Interests and Funding: The authors declare no conflicts of interest. This research received no specific grant from public, commercial, or not-for-profit funding agencies.

REFERENCES

1. Mueller N, Westerby M, Nieuwenhuijsen M. Health impact assessments of shipping and port-sourced air pollution on a global scale: A scoping literature review. Environmental research, 2023 Jan. 216 (Pt 1), 114460. https://doi.org/10.1016/j.envres.2022.114460
2. Schnurr REJ, Walker TR. Marine transportation and energy use. Environmental Earth Sciences, 2019,29:1-7. https://doi.org/10.1016/b978-0-12-409548-9.09270-8.
3. European Commission. Reducing emissions from the shipping sector. 2023. [Retrieved November 12, 2023]. Available from: https://climate.ec.europa.eu/eu-action/transport/reducing-emissions-shipping-sector_en.
4. United Nations Conference on Trade and Development (UNCTAD). Review of Maritime Transport 2023: Towards a green and just transition. 2023 Sep. [Retrieved November 12, 2023]. Available from: https://unctad.org/publication/review-maritime-transport-2023.
5. International Maritime Organization (IMO). Fourth Greenhouse Gas Study 2020. 2020. [Retrieved November, 12, 2023]. Available from: https://www.imo.org/en/ourwork/Environment/Pages/Fourth-IMO-Greenhouse-Gas-Study-2020.aspx
6. Basan F, Fischer J, Kühnel D. Soundscapes in the German Baltic Sea Before and During the Covid-19 Pandemic. Frontiers in Marine Science, 2021 Jun. 8. https://doi.org/10.3389/fmars.2021.689860
7. Song C, He J, Wu L, et al. Health burden attributable to ambient PM2.5 in China. Environmental Pollution, 2017 Apr. 223: 575–86. https://doi.org/10.1016/j.envpol.2017.01.060
8. Endresen O, Sorgard E, Behrens HL, Brett PO, Isaksen IS. A historical reconstruction of ships' fuel consumption and emissions. Journal of Geophysical Research Atmospheres, 2007 Jun. 112(12): 17. https://doi.org/10.1029/2006JD007630
9. Gibney E. Coronavirus lockdowns have changed the way Earth moves. Nature, 2020 Mar. 580(7802): 176–77. https://doi.org/10.1038/d41586-020-00965-x
10. Peters MD, Godfrey C, McInerney P, et al. Chapter 11: Scoping Reviews (2020 version). JBI Manual for Evidence Synthesis [Internet]. JBI, 2020. [cited 2023 Mar 27]. Available from: https://synthesismanual.jbi.global.
11. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine, 2018 169(7):467-473. doi: 10.7326/M18-0850.
12. Open Science Framework (OSF). Available from: https://osf.io/cknhv/. [Accessed July 30, 2024].
13. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Systemaic Review, 2016 Dec. 5(1): 210. https://doi.org/10.1186/s13643-016-0384-4.
14. Abous H, Hamiche ME, El Merouani M. The impact of Covid-19 on the port environment: the case of tanger med container port, Morocco. E3S Web Conf; 2021,234:00025. https://doi.org/10.1051/e3sconf/202123400025.
15. Chen L, Li J, Pang X, et al. Impact of COVID-19 Lockdown on Air Pollutants in a Coastal Area of the Yangtze River Delta, China, Measured by a Low-Cost Sensor Package. Atmosphere. 2021 Mar 6;12(3):345.
16. Durán-Grados V, Amado-Sánchez Y, Calderay-Cayetano F, et al. Calculating a Drop in Carbon Emis-

sions in the Strait of Gibraltar (Spain) from Domestic Shipping Traffic Caused by the COVID-19 Crisis. Sustainability. 2020 Dec 11;12(24):10368.
17. Gu B, Liu J. COVID-19 pandemic, port congestion, and air quality: Evidence from China. Ocean & Coastal Management. 2023 Mar;235:106497.
18. He Z, Lam JSL, Liang M. Impact of Disruption on Ship Emissions in Port: Case of Pandemic in Long Beach. Sustainability [Internet]. 2023 Jan 1 [cited 2023 Dec 8];15(9):7215. Available from: https://www.mdpi.com/2071-1050/15/9/7215
19. Ju Y, Hargreaves CA. The impact of shipping CO2 emissions from marine traffic in Western Singapore Straits during COVID-19. Science of The Total Environment. 2021 Oct;789:148063.
20. Mocerino L, Quaranta F. How emissions from cruise ships in the port of Naples changed in the COVID-19 lockdown period. Proceedings of the Institution of Mechanical Engineers Part M, Journal of engineering for the maritime environment/Proceedings of the Institution of Mechanical Engineers Proceedings part M, Journal of engineering for the maritime environment. 2021 Jun 28;236(1):125–30.
21. Mujal-Colilles A, Guarasa JN, Fonollosa J, Llull T, Castells-Sanabra M. COVID-19 impact on maritime traffic and corresponding pollutant emissions. The case of the Port of Barcelona. Journal of Environmental Management. 2022 May;310:114787.
22. Sarra SR, Mülfarth RCK. Impactos das atividades portuárias sobre a poluição atmosférica na cidade de Santos (Brasil). CONJ [Internet].2022,22(2):1-14.
23. Shi K, Weng J. Impacts of the COVID-19 epidemic on merchant ship activity and pollution emissions in Shanghai port waters. Sci Total Environ. 2021 Oct 10;790:148198. https://doi.org/10.1016/j.scitotenv.2021.148198.
24. Zabbey N, Sam K, Newsom CA, Nyiaghan PB. The COVID-19 lockdown: An opportunity for conducting an air quality baseline in Port Harcourt, Nigeria. The Extractive Industries and Society. 2021 Mar;8(1):244–56.
25. Badurina, P.; Cukrov, M.; Dundović, Č. Contribution to the implementation of "Green Port" concept in Croatian seaports. Pomorstvo 2017, 31:10–17.

Sample Size for Agreement Studies on Quantitative Variables

Bruno Mario Cesana(1), *Paolo Antonelli*(2)

(1) Associate Professor of Medical Statistics. Retired from the Section of Medical Statistics and Biometry, Department of Molecular and Translational Medicine, Faculty of Medicine and Surgery, University of Brescia, V. le Europa 11, 25123 Brescia, Italy and Department of Clinical Sciences and Community Health, Unit of Medical Statistics, Biometry and Bioinformatics "Giulio A. Maccacaro" Faculty of Medicine and Surgery, University of Milan, Via Celoria 22, 20133, Milan, Italy.
(2) Retired as a Professor of Calculus of Probabilities, Statistics and Operative Research at the State Industrial Technical Institute (ITIS) Benedetto Castelli, Brescia, Via Antonio Cantore, 9, 25128 Brescia, Italy.

CORRESPONDING AUTHOR: Bruno Mario Cesana Phone. +39 02503 - 20854 Fax 02503 – 20855. Email: brnmrcesana@gmail.com

SUMMARY

We reviewed the statistical assessments of the agreement between two measurement methods of continuous variables together with their recent contributions about the sample size calculation based on the "two one side t-tests (TOST) extensions to the individual equivalence. We generalized a restricted null hypothesis that constitutes a particular case in finding the supremum of the probability of rejecting the equivalence under the null hypothesis (H_0) and which, obviously, limits its applicability.
Particularly, we devise and propose an exact procedure for calculating the sample sizes for individual equivalence, as an expression of the agreement between two measurement methods, by using a size a test (that is, with adequate control of Type I error), based on the non-central bivariate t distribution with correlation equal to 1 and to the related functions for calculating a and 1-b probabilities.
Furthermore, our devised procedure allows to calculate the sample sizes by choosing between two most suitable formulations of the global parameters space of the null and alternative hypotheses; indeed, they are based on the portion of the distribution of the differences between the two measurement methods or on appropriately chosen agreement thresholds.
Thereafter, we compared our theoretical results with the recently published proposals of the sample size calculation for the Bland and Altman agreement analysis by means also of simulation studies.
Finally, a program written in the open-source R language to perform sample size calculations according to our procedure is available upon request.

Keywords: Measurement Methods comparison; Quantitative variables; Bland-Altman analysis; Sample size calculation; Individual equivalence.

INTRODUCTION

Let's define the statistical model of an agreement study between two measurement methods of quantitative variables without replicates, according to the usual model of study carried out in clinical and, above all, laboratory settings. It has to be noted that in the context of an agreement study the two measurement methods under comparison are expected to be of equivalent precision. Particularly, there is an "Old measurement method" that can be defined as "Standard" (but without the connotation of a "Gold Standard") and a "New measurement method" that can be defined as "Experimental" or "Test" (thereafter, Test for simplicity). In addition, the Test method has some advantages over the current Standard (lower cost, greater simplicity of execution, for example), but to replace the "Old/ Standard measurement method" it must be proved to be "essentially equivalent" in the context of an agreement study.

Let's assume that the variable X, Gaussian distributed $[X \sim G(\mu_X, \sigma_X^2)]$, represents the values of the Standard and that the variable Y, Gaussian distributed $[Y \sim G\left(\mu_Y, \sigma_Y^2\right)]$, represents the values of the Test.

Thus, a measurement value of the Standard can be expressed as: $X_i = \beta_1\mu_i + \xi_1 + \varepsilon_{iX}$ and a measurement value of the Test can be expressed as: $Y_i = \beta_2\mu_i + \xi_2 + \varepsilon_{iY}$. From the above definitions, it is

DOI: 10.54103/2282-0930/23479
Accepted: 23th May 2024

assumed that each value is obtained by the sum of the true value (μ_i with i = 1, ..., n) of the i-th subject/laboratory sample multiplied by a fixed constant (b_1 or b_2), of the effect of the measurement method (ξ_j, with j = 1, 2 for the Standard and the Test, respectively), and of the measurement error ξ_{ij}, considered individually and independently distributed for the i-th subject/laboratory sample. This error component, is assumed (as a usual assumption) to be Gaussian and independently distributed: $\left[\varepsilon_X \sim G(0,\sigma^2_{\varepsilon X})\right]$ and $\left[\varepsilon_Y \sim G(0,\sigma^2_{\varepsilon Y})\right]$, respectively.

Furthermore, it has to be outlined that $b_1 \neq 1$ and $b_2 \neq 1$ cause the occurrence of a proportional error which can be of different magnitude for the two measurement methods under comparison. In addition, if $\xi_1 \neq 0$ and $\xi_2 \neq 0$, there will be a systematic error for the two measurement methods under comparison. Indeed, the paired differences between measurement methods in perfect agreement lie on the bisecting line of the first cartesian plane with intercept equal to 0 and slope equal to 1.

Otherwise, the systematic and proportional error between the two measurement methods will be considered when, in the agreement analysis the differences will be regressed on their means according to Bland and Altman [4,5]. Obviously, the difference between the population means (ξ_1 and ξ_2) of the two measurement methods can be the source of a possible "systematic error".

Taking into account the agreement between the two measurement methods under comparison, we focus on their paired sampling differences:

$$\begin{aligned} D_i &= X_i - Y_i = \beta_1\mu_i + \xi_1 + \varepsilon_{iX} - \left(\beta_2\mu_i + \xi_2 + \varepsilon_{iY}\right) \\ &= \beta_1\mu_i - \beta_2\mu_i + \xi_1 - \xi_2 + \varepsilon_X - \varepsilon_Y = \xi_1 - \xi_2 + \varepsilon_X - \varepsilon_Y \end{aligned}$$

These differences (D_i) are given by the difference between the values of the two measurement methods and of their measurement errors, since, under the agreement assumption, the multiplicative constant of the true values are assumed to be equal ($b_1 = \beta_2 = \beta$, even if they are not equal to 1) and the true values (μ_i), equal components of the two above reported expressions, are removed together with the biological variability of each subject/laboratory sample.

Finally, D_i, being the difference of two independent Gaussian distributed variables with measurement error, is Gaussian distributed with mean μ_D, and variance $\sigma^2_D = \sigma^2_{\varepsilon X} + \sigma^2_{\varepsilon Y}$, both unknown. However, under the agreement situation, μ_D is expected to be almost zero and $\sigma^2_D \approx 2\sigma^2_{\varepsilon X}$ or $\approx 2\sigma^2_{\varepsilon Y}$, since it is expected that $\sigma^2_{\varepsilon X} \approx \sigma^2_{\varepsilon Y}$ with correlation equal to zero. Thereafter, μ_D and σ_D will be indicated as μ and σ for semplicity.

It should be emphasized that agreement studies carried out in clinical settings are not always preceded by the pertinent assessments of accuracy, precision, reliability, repeatability and reproducibility of the new (Test) measurement method as practically always happens in laboratory environments where there is a greater knowledge of the measurement aspects and a stricter adherence to the pertinent guidelines. Therefore, it is possible that a new clinical measurement method is compared with the current standard without having satisfactory metric properties.

Statistical analysis and Sample size for agreement studies

It is widely recognized that biomedical researches have to be adequately powered in order to have a satisfactorily high probability to achieve their (primary) objectives.

Of course, also agreement studies between two (or more) measurement methods need to be adequately powered, although there does not seem to be adequate attention to this aspect, as reported by a recent review by Han et al. [1]. Particularly, Han et al. [3] wrote that: "only 27 studies out of 82 (33%) gave justification for their sample size".

Furthermore, also Kottner et al. [4] and Gerke et al. [5] have pointed out that formal justifications for sample size were rarely reported in agreement studies.

We are interested in sample size calculations for agreement studies on quantitative variables carried out according to the Bland and Altman approach [1,2] (thereafter B&A, for simplicity).

Indeed, Han et al. [3] reported that agreement studies on continuous variables are very frequently carried out by calculating the limits of agreement (LoAs) according to B&A's method [1,2] that "enables us to separate systematic and random error, which r (the correlation coefficient: authors' note), combines into a single measure" as B&A wrote [6].

However, it has to be stressed that researchers often consider and use the B&A's procedure [1,2] simply as a "graphical approach" without paying due attention to its assumptions and conditions which have to be fulfilled in order to draw valid conclusions. Particularly, the Gaussian distribution of the differences, the equality of the variances of the two measurement methods, the absence of relevant systematic bias and, above all, the absence of relevant proportional bias, as Taffè [7] has recently reiterated. Furthermore, it has also to be said that, unfortunately, these assumptions are unlikely to hold in practice.

It has to be stressed that the sample size calculation for agreement studies has progressively been moved to the sample size for equivalence studies. Furthermore, from equivalence studies focussed on the equivalence of means (ABE: Average BioEquivalence), the recent methodological contributes deal with the Population BioEquivalence (PBE) aspect in which the agreement limits must include a relevant part of the central population of the differences between two measurement methods, and, finally with the individual bioequivalence (IBE) where the assessment is made on the differences between values measured on the same unit and, consequently, paired. Particularly, the usual 95% limits of agreement (LoA) encompass the 2.5th percentile and 97.5th percentile of the distribution of the paired differences between the values recorded by the measurement methods on the same specimen/subject. Accordingly, sample sizes procedures have been proposed.

Therefore, we considered outdated the proposed sample sizes calculations based on the precision of the 95% Confidence Interval (CI) of the LoAs according to a generic recommendation given by Bland on his website [8]: "I usually recommend 100 as a good sample size, which gives a 95% CI about ± 0.34s, being "s" the standard deviation of the differences between the measurements of the two methods. A sample of 200 subjects is even better, giving a 95%CI about ± 0.24s. As with all estimation, to determine the appropriate sample size the researchers must decide what accuracy is required."

Moreover, Bland's suggestion [8] cannot be considered completely shareable owing to the fact that the required precision of the LoAs can be obtained only about the 50% of the cases as firstly highlighted by Cesana and Antonelli [9] and recognized also by Lu et al. [10].

Likewise, we do not consider the sample sizes calculations approaches by Shieh [11] and Shieh [12] with the exact centile calculation, shown by Carkeet [13] and by Carkeet and Goh [14], for having an adequate probability, defined as "assurance probability", of obtaining 95%CI intervals width of the LoA less than a required width. Similarly, we do not consider the sample size calculation proposed by Jan and Shieh [15].

Bland-Altman's method and its sample size calculation

We focus our attention on the sample size calculation proposed by Shieh [16] together with the TOST procedures for assessing the agreement proposed by Liu and Chow [17], Lin et al. [18], and Lu et al. [10], considered in Shieh's paper [16], together with the B&A's approach [1,2] based on the approximate confidence intervals of normal centiles.

Considering that "to establish the agreement between two methods, the central portion of the distribution of paired differences needs to be within a close range around zero", Shieh [16] calculated the supremum of the Type I error (α) under a null hypothesis of no equivalence attained "when the two centiles coincide with the boundary values $\left\{\hat{\theta}_{1-p};\hat{\theta}_{p}\right\}=\left\{-\Delta;\Delta\right\}$". This sentence leads to the simultaneous equality of $\hat{\theta}_{1-p}$ with $-\Delta$ and $\hat{\theta}_{p}$ with Δ.

Rather curiously, also in the Krishnamoorthy and Mathew's book [19] (on page 35, equation 2.3.11) there is a similar unproved sentence "note that the supremum in the above equation is attained at $L_l=\mu-z_{(1+p)/2}\sigma$ and $L_u=\mu+z_{(1+p)/2}\sigma$" where the (1+p)/2 subscript corresponds to the population proportion, defined p in our notation. It has to be stressed that the affirmation of Krishnamoorthy and Mathew [19], followed by Shieh [16], is valid only in a particular subset of H_0 as we will demonstrate and, consequently, it has to be considered unsuitable because it is not general.

Moreover, consequently to his definition of the supremum under H_0, Shieh [16] calculated the critical value $\gamma_{1-\alpha}$ such that the statistical test is of size α. Then, the statistical agreement test rejects the null hypothesis if $\gamma_{1-\alpha} < T_L$ and $T_U < -\gamma_{1-\alpha}$ where $T_L=\left(\bar{D}+\Delta\right)\Big/\sqrt{\left(S^2/n\right)}$ and $T_U=\left(\bar{D}-\Delta\right)\Big/\sqrt{\left(S^2/n\right)}$ where $\bar{D}$ and S^2 are the sample mean and the variance of the differences, respectively, n is the sample size and Δ is the equivalence threshold.

Furthermore, consistently with his calculated critical value $\gamma_{1-\alpha}$, Shieh [16] claimed that the Type I errors of the considered TOST approaches by Liu and Chow [17], B&A [2], Lin et al. [18] and Lu et al. [10] turned out to be too much conservative.

Aim of the paper

Aim of our paper is to show the appropriate sample size calculation approach for agreement studies carried out according to the "individual equivalence" model. It has to be noted that the appropriate approach must consider all the possible pair (μ;σ) under H_0 and H_A and, consequently, it should not be limited to particular cases.

Indeed, we have found that Shieh's proposal [16] is valid only in a particular context and, consequently we propose an exact general procedure for calculating the sample size for individual equivalence by using a size α test (that is, with adequate control of Type I error), as defined by Casella and Berger [20], based on the non-central bivariate t distribution, with correlation equal to 1 and Owen's Q functions [21] for calculating α and 1-β probabilities.

Furthermore, we consider two different sample size calculation approaches: the first is based on the population proportion (p) or on its corresponding normal quantile (z_p), and the second is based on the agreement threshold (Δ).

Particularly, keeping fixed Δ, it is possible to hypothesize a population proportion under H_A greater than the population proportion under H_0 ($p_A > p_0$) or, equivalently, in the terms of the quantiles $z_{pA} > z_{p0}$. Vice versa, keeping fixed the population proportion (p) or the normal quantile z_p, it is possible to hypothesize, under H_A, a narrower interval ($-\Delta_A$, Δ_A) than under H_0 ($-\Delta_0$, Δ_0).

Furthermore, we compare our results with those coming from the recent proposal of Shieh [16] and in particular with those from Liu and Chow [17] procedure.

Finally, we clarify whether the TOST procedure applied to the individual equivalence is of size α or of level α, according to the definition of Casella and Berger [20].

The Methods section details the theoretical approach of the sample size calculation. Even if the more technical aspects have been moved to appropriate appendices so as not to interrupt the flow of presentation and of the reasoning, the following topics will be considered: (i) Null (H_0) and alternative (H_A) hypotheses together with their parameter space; (ii) Outlines of the proposed procedure; (iii) The probability of rejecting H_0 and its calculation. Then, the paragraph "Sample size calculation" shows: (A) the determination of the non-centrality parameters under H_0 and H_A; (B) the alternative hypothesis according to two proposed different approaches (Case 1: model with fixed population proportion p_0. Case 2: model with fixed agreement threshold Δ); and (C) the Sample size calculation procedure.

The Results section shows the "Tables of the sample size" for the main common and sensible scenarios of agreement studies, the "Comparisons between the sample sizes calculated with p fixed (Case 1) and Δ fixed (Case 2)" and "a particular approach: the sample size calculation under two simple hypotheses". Then, there are the results of simulation studies focussing on: (A) Check of the fulfilment of the nominal significance level α, and (B) Check of the fulfilment of the nominal power (1 - β). A paragraph about the "comparisons among the different considered methods: our new AC procedure, Shieh [16], Liu and Chow [17], Bland and Altman [2], Lin et al. [18] and Lu et al. [10] and a paragraph "Check of the fulfilment of the nominal significance level α of the various methods" follow.

The two paragraphs "Considerations about the canonical null hypothesis H_0 and Shieh's approach [16]" and "Considerations about Liu and Chow's TOST procedure [17] and our AC procedure" precede the Discussion section that concludes the paper.

METHODS

Let's denote the 100p-th centile of the Gaussian distribution G(μ,σ²) as $\theta_p = \mu + z_p\sigma$, where z_p is the 100p-th centile of the standard Gaussian distribution G(0, 1); it has to be noted that we use the lowercase "p" notation to define the centile of a distribution.

Null (H_0) and Alternative (H_A) Hypotheses together with their Parameter Space

To establish the agreement between two measurement methods, a relevant central portion (defined as a capital "P": P = 2p-1) of the distribution of the paired differences needs to be within a narrow interval (say, -Δ, Δ) around zero.

Thus, the statistical test is formulated as an equivalence test with the null (H_0) and the alternative (H_A) hypotheses given by:

$$H_0 : \left(\theta_{1-p} \le -\Delta\right) \vee \left(\theta_p \ge \Delta\right)$$

$$\text{vs. } H_A : \left(-\Delta < \theta_{1-p}\right) \wedge \left(\theta_p < \Delta\right) \qquad \text{Formula 1}$$

Otherwise, being $\theta_{1-p} = \mu - z_p\sigma$, $\theta_p = \mu + z_p\sigma$ with p >0.5, $\mu \in R$ and $\sigma > 0$,

$$H_0 : \left(\mu - z_p\sigma \le -\Delta\right) \vee \left(\mu + z_p\sigma \ge \Delta\right) \text{ vs.}$$

$$H_A : \left(-\Delta < \mu - z_p\sigma\right) \wedge \left(\mu + z_p\sigma < \Delta\right), \text{ respectively.}$$

The null (H_0) and the alternative (H_A) hypotheses depend on μ and σ, being fixed the other parameters (p and Δ). These hypotheses have an adequate geometrical representation in the R² space, by placing μ on the horizontal X axis and σ on the vertical Y axis as shown in Figure 1 (Panel A and Panel B).

The alternative hypothesis H_A can be reported as:

$$-\Delta < \mu < \Delta \quad \text{and ,}$$

$$\sigma < \left(\Delta - \mu\right) / z_p \wedge \sigma < \left(\Delta + \mu\right) / z_p \wedge \sigma > 0$$

ultimately as:

$$H_A = \left\{ \begin{matrix} \left(\mu;\sigma\right) : \left(-\Delta < \mu < \Delta\right) \wedge \\ \left[0 < \sigma < \left(\Delta - |\mu|\right) / z_p\right] \end{matrix} \right\} \qquad \text{Formula 2}$$

Thus, the H_A space corresponds to the points (μ; σ) inside the ABC triangle, shown in Figure 1 (Panel B), determined by the lines: σ = 0; $\sigma = \left(\Delta - \mu\right) / z_p; \sigma = \left(\Delta + \mu\right) / z_p$

The null hypothesis H_0 is the complementary set of H_A and, therefore, is simply the set of points (μ;σ) of the positive half-plane of the ordinates (σ > 0) not inside the triangle ABC.

Formally, H_0 is the union of two subspaces: $H_0 = H_{0A} \cup H_{0B}$, where:

$H_{0A} = \left\{\left(\mu;\sigma\right) : \left(\mu - z_p\sigma \le -\Delta\right)\right\} = \left\{\left(\mu;\sigma\right) : \left(-\infty < \mu < \infty\right) \wedge \sigma \ge \left(\Delta + \mu\right) / z_p > 0\right\}$ is the region on the left of the half-line on which the side AC lies, and

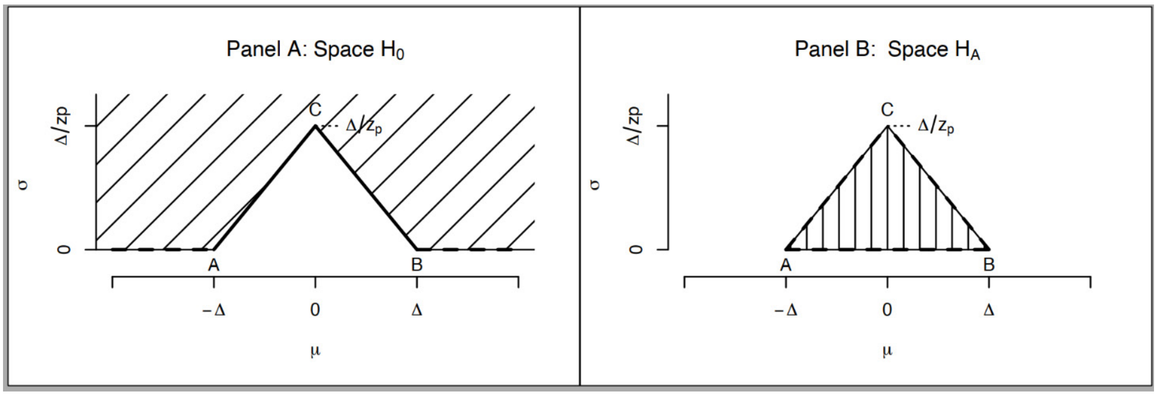
$H_{0B} = \left\{\left(\mu;\sigma\right) : \left(\mu + z_p\sigma \ge \Delta\right)\right\} = \left\{\left(\mu;\sigma\right) : \left(-\infty < \mu < \infty\right) \wedge \sigma \ge \left(\Delta - \mu\right) / z_p > 0\right\}$ is the region on the right of the half-line on which the side BC lies.

Thereafter, H_0 can be briefly defined as:

$$H_0 = \left\{ \begin{matrix} \left(\mu;\sigma\right) : \left(-\infty < \mu < +\infty\right) \wedge \\ \sigma \ge \left(\Delta - |\mu|\right) / z_p > 0 \end{matrix} \right\} \qquad \text{Formula 3}$$

Thus, the H_0 space is represented by the positive half-plane of the Y axis with the exclusion of the inner points of the ABC triangle, with vertices A = (-Δ; 0),

Figure 1. Space of the null hypothesis (H_0, Panel A) and space of the alternative hypothesis (H_A, Panel B)



B = (Δ; 0), and C = (0;σ = Δ/z_p), shown in Figure 1 (Panel A).

It should be noted that the alternative hypothesis H_A specifies that there is at least P = 2p -1 central portion of the distribution of the paired differences in the range (−Δ, Δ), while the null hypothesis H_0 specifies that the central portion P included in the range (−Δ, Δ), is less than 2p-1.

Outlines of the proposed procedure

Following Shieh [16], "a natural rejection region to assess", at a significance level α, a central P portion $(\theta_{1-p};\theta_p)$ of a Gaussian population of the differences D~G(μ;σ) with the p-centile "p" equal to: p = (P + 1)/2 is given by:

$$E=\left\{-\Delta<\hat{\theta}_{1-p}\wedge\hat{\theta}_p<\Delta\right\} \qquad \text{Formula 4}$$

with $\hat{\theta}_{1-p}=\bar{D}-k_{1-\alpha}S/\sqrt{n}$ and $\hat{\theta}_p=\bar{D}+k_{1-\alpha}S/\sqrt{n}$, being $k_{1-\alpha}$ an opportune constant to be determined to control the Type I error probability; furthermore, "$\bar{D}$" and "S" are the sample mean and the sample standard deviation of the distribution of the differences, respectively.

Denoting with Pr{E} the probability of rejecting H_0, we obtain: $Pr\{E\}=\Pr\left\{\left(-\Delta<\hat{\theta}_{1-p}\right)\wedge\left(\hat{\theta}_p<\Delta\right)\right\}$

Our sample size calculation procedure is based on the two following main steps.

a. In order to keep the Type I error always less than or equal to the prefixed α for any pair of values (μ;σ) within the H_0 space, the $k_{1-\alpha}$ coefficient has to be determined so that the superior of the probability of the event E is equal to α: $\sup_{H_0}Pr\{E\}=\alpha$. The superior is calculated in order that the Type I error is always ≤ α.
We will demonstrate that, under H_0, the Type I error reaches its supremum value (α) in the two points $(\mu;\sigma)=\left[\mu;\left(\Delta-|\mu|\right)/z_p\right]$ for $\mu\to-\Delta^+$ and for $\mu\to\Delta^-$ corresponding to the points A and B of the ABC triangle (Figure 1, Panel A), instead of the point $(\mu;\sigma)=\left(0;\Delta/z_p\right)$, claimed by Shieh [16], corresponding to the vertex C of the ABC triangle (Figure 1, Panel B).
b. The sample size "n" is calculated so that the $\inf_{H_A}Pr\{E\}=1-\beta$. Thus, the power of the test is always not inferior to the prefixed threshold of 1 - β for any pair of values (μ;σ) in the H_A space. We will demonstrate that, under H_A, the power attains its lower extremum in the point $(\mu;\sigma)=\left(0;\Delta/z_p\right)$ corresponding to the vertex C of the ABC triangle (Figure 1, Panel B), in agreement with Shieh [16].

The sample size calculation is based on the non-central bivariate t distribution with correlation equal to 1 and Owen's Q formulation, together with some related theorems [21].

The probability of rejecting H_0: Pr{E}

The event $E=\left\{-\Delta<\hat{\theta}_{1-p}\wedge\hat{\theta}_p<\Delta\right\}$ after some algebraic steps, can be written as:

$$E=\left\{\begin{matrix}\left(\bar{D}+\Delta\right)/\left(S/\sqrt{n}\right)>k_{1-\alpha}\wedge\\ \left(\bar{D}-\Delta\right)/\left(S/\sqrt{n}\right)<-k_{1-\alpha}\end{matrix}\right\} \qquad \text{Formula 5}$$

Owen [21] (page 437) wrote that the statistics:

$T_L=\left(\bar{D}+\Delta\right)/\left(S/\sqrt{n}\right)$ and $T_U=\left(\bar{D}-\Delta\right)/\left(S/\sqrt{n}\right)$ are distributed as non-central Student's t distributions, with ν = n – 1 degrees of freedom and non-centrality parameter τ given by: τ_L=(μ+Δ)/(σ/√n) and τ_u=(μ-Δ)/(σ/√n), respectively. Our demonstration of Owen's affirmation [21] is given in the Appendix A.

Furthermore, if these statistics are jointly considered, they are distributed as a non-central bivariate t with correlation equal to 1, according to Owen [21] who refined the definition of a multivariate t-distribution given by Dunnett and Sobel [22].

Thus, the probability of the event E is:

$$\Pr\{E\} = \Pr\{(T_L > k_{1-\alpha}) \wedge (T_U < -k_{1-\alpha})\} = \\ = \Pr\ \{[t_{\nu,\tau_L} > k_{1-\alpha}] \wedge [t_{\nu,\tau_U} < -k_{1-\alpha}]\} \qquad \text{Formula 6}$$

where $[t_{\nu,\tau L};t_{\nu,\tau U}]$ is a non-central bivariate t-distribution with degrees of freedom ν, non-centrality parameters τ_L and τ_U, and correlation equal to 1.

Particularly, Pr{E} can be calculated using statistical packages that implement the non-central bivariate t distribution such as Owen's package "OwenQ" [23] that implements also the well-known Owen's Q functions [21] or "PowerTOST" package [24] or using the package "mvtnorm" from Genz et al. [25] with a non-deterministic procedure.

From Owen's formulas [21] for calculating Pr{E}, it is possible to note the following two useful properties of Pr{E}:

a. Pr{E} depends on the non-centrality parameters τ_L and τ_U. Particularly, Pr{E} increases when τ_L increases and when τ_U decreases, as it is possible to verify by calculating the area of the cumulative non-central bivariate t density function. For example, with n = 134 and $k_{1-\alpha}$ =17, when τ_L = 15 and τ_U = -15, Pr{E} = 0.00752; moreover, when τ_L = 21 and τ_U = -15, Pr{E} = 0.0857. Finally, when τ_L = 21 and τ_U = -21, Pr{E} = 0.99437.
b. $\Pr\{E \mid \tau_L, \tau_U\} = \Pr\{E \mid -\tau_U, -\tau_L\}$. That is, Pr(E) value does not change even when $\tau_L = -\tau_U$ and $\tau_U = -\tau_L$.

Calculation of $\sup_{H_0} \Pr\{E\}$

Theorem: $\sup_{H_0} \Pr\{E\}$, under H_0, is attained as a limit when the generic point (μ;σ) tends to A (-Δ; 0) on the AC side or tends to B (Δ; 0) on the BC side of the ABC triangle shown in Figure 1 (Panel A).

UnderH$_0$, the conditions are: $(-\infty < \mu < +\infty)$ and σ ≥ $(\Delta - |\mu|)/z_p > 0$.

It is well evident that the critical points lie on the boundary of the H_0 space; particularly, on the sides AC $(\mu_0 - z_p\sigma_0 = -\Delta)$ or CB $(\mu_0 + z_p\sigma_0 = \Delta)$ of the ABC triangle excluded its A and B points since, obviously, σ > 0.

Considering the points on the side AC, the conditions are:

$$-\Delta < \mu \le 0 \text{ and } \sigma = (\Delta - |\mu|)/z_p = (\Delta + \mu)/z_p.$$

Consequently, the two non-centrality parameters become:

$$\tau_L = \frac{\mu + \Delta}{\sigma/\sqrt{n}} = \frac{\mu + \Delta}{\Delta - |\mu|} \cdot z_p\sqrt{n}$$
$$= \frac{\mu + \Delta}{\Delta + \mu} \cdot z_p\sqrt{n} = z_p\sqrt{n} \text{ and}$$
$$\tau_U = \frac{\mu - \Delta}{\sigma/\sqrt{n}} = \frac{\mu - \Delta}{\Delta - |\mu|} \cdot z_p\sqrt{n} = \frac{\mu - \Delta}{\Delta + \mu} \cdot z_p\sqrt{n}$$

Therefore, on the side AC of the ABC triangle, τ_L is a constant, while τ_U is a branch of a hyperbola; particularly, we obtain the following limits:

$$\lim_{\mu \to -\Delta^+} \tau_L = z_p\sqrt{n} \text{ and } \lim_{\mu \to -\Delta^+} \tau_U = -\infty .$$

Furthermore, for the side BC of the ABC triangle the conditions are:

$$0 \le \mu < \Delta \text{ and } \sigma = (\Delta - |\mu|)/z_p = (\Delta - \mu)/z_p$$

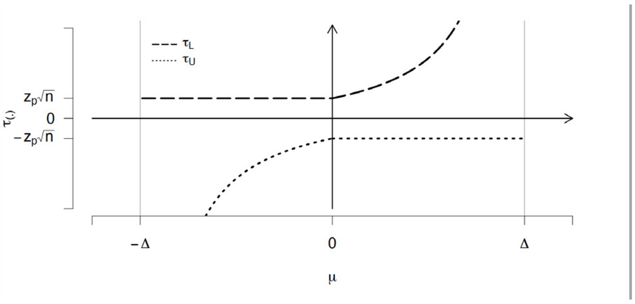
Consequently, the two non-centrality parameters become:

$$\tau_L = \frac{\mu + \Delta}{\sigma/\sqrt{n}} = \frac{\mu + \Delta}{\Delta - \mu} \cdot z_p\sqrt{n} \text{ and}$$
$$\tau_U = \frac{\mu - \Delta}{\sigma/\sqrt{n}} = \frac{\mu - \Delta}{\Delta - \mu} \cdot z_p\sqrt{n} = -z_p\sqrt{n}$$

Therefore, on the side BC of the ABC triangle, τ_L is a branch of a hyperbola, while τ_U is a constant; particularly, we obtain the following limits: $\lim_{\mu \to \Delta^-} \tau_L = +\infty$ and $\lim_{\mu \to -\Delta^-} \tau_U = -z_p\sqrt{n}$.

Figure 2 shows the graphs of τ_L (long-dashed line) and τ_U (dashed line) functions against μ. It is possible to see that the two functions reach their vertical asymptotes (+∞; -∞) when μ tends to +Δ or to -Δ, respectively.

Figure 2. Graphs of the τ_L (long-dashed line) and τ_U (dashed line) functions against μ



By recalling the properties of Pr{E}, $\sup_{H_0} \Pr\{E\}$ is obtained when $\mu \to +\Delta^-$ and $\sigma = \lim_{\mu \to \Delta^-} (\Delta - |\mu|)/z_p = 0^+$, since τ_L increases to +∞ and τ_U is constant or when $\mu \to -\Delta^+$ and $\sigma = \lim_{\mu \to -\Delta^+} (\Delta - |\mu|)/z_p = 0^+$, since τ_L is constant and

τ_U decreases to -∞. This occurs on the boundary of the H_0 space at the points B and A of the ABC triangle shown in Figure 1 (Panel A).

Thus, the superior under H_0 of the Pr{E} is given by:

$$\sup_{H_0}\Pr\{E\}=\sup_{H_0}\Pr\left\{\begin{matrix}(T_L>k_{1-\alpha})\wedge(T_U<-k_{1-\alpha})\\ |\tau_L=+\infty;\tau_U=-z_p\sqrt{n}\end{matrix}\right\}=$$

Formula 7

$$=\Pr\left\{\begin{matrix}(t_{n-1,\tau_L}>k_{1-\alpha})\wedge(t_{n-1,\tau_U}<-k_{1-\alpha})\\ |\tau_L=+\infty;\tau_U=-z_p\sqrt{n}\end{matrix}\right\}$$

at the point B. It has to be noted that $[t_{v,\tau L}; t_{v,\tau U}]$ is a non-central bivariate t with correlation equal to 1.

In the case of the point A of the ABC triangle, the conditioning is given by: $\tau_L = z_p\sqrt{n}$; $\tau_U = -\infty$.

It is worthwhile to stress that the superior of the Pr{E} under H_0 is equal in the points A and B of the ABC triangle, according to the previously reported property b of Pr{E}.

This conclusion disagrees with Shieh's affirmation [16] that the upper extremum is obtained when $(\theta_{1-p}=-\Delta)\wedge(\theta_p=\Delta)$, or, equivalently, when $(\mu=0)\wedge(\sigma=\Delta/z_p)$ at the coordinates of the vertex C of the ABC triangle.

We will demonstrate in the following that Shieh's affirmation [16] is valid only under a condition more restrictive than that foreseen by the canonical H_0.

According to a geometrical approach, it has to be noted that starting from the vertex C ($\mu = 0$ and $\sigma = \Delta/z_p$) of the ABC triangle, and going down on the side AC the probability Pr{E} increases owing to the fact that τ_L is constant and τ_U decreases tending to -∞. Furthermore, going down on the side BC, Pr{E} increases since τ_L increases tending to +∞ and τ_U is constant. So, the point C is not the place of the supremum of Pr{E] as Shieh affirmed [16].

In fact, the point C of the ABC triangle has coordinates: $\tau_L=z_p\sqrt{n}$ and $\tau_U=-z_p\sqrt{n}$ very far from the points A or B where Pr{E} reaches its supremum value (Figure 1, Panel A).

Consequently, we have appropriately modified Shieh's [16] sample size calculation and we will give the correct formulation for the general canonical case.

Calculation of $\inf_{H_A}\Pr\{E\}$

Theorem: $\inf_{H_A}\Pr\{E\}$ *, under* H_A*, corresponds to the point C (*$\mu = 0$*,* $\sigma=\Delta/z_p$*).*

Under H_A, the conditions are: $-\Delta<\mu<\Delta$ and $0<\sigma<(\Delta-|\mu|)/z_p$ corresponding to the inner points of the ABC triangle.

Also in this case, the points at which Pr{E} attains its infimum are located on the edge of the H_A space, particularly on the sides AC and BC of the ABC triangle.

Considering the points on the side AC and BC, we have: $-\Delta<\mu<\Delta$ and $\sigma=(\Delta-|\mu|)/z_p$; thus, the two non-centrality parameters are:

$$\tau_L=(\mu+\Delta)/(\sigma/\sqrt{n})$$
$$=(\mu+\Delta)/(\Delta-|\mu|)/z_p\sqrt{n}$$

and

$$\tau_U=(\mu-\Delta)/(\sigma/\sqrt{n})$$
$$=(\mu-\Delta)/(\Delta-|\mu|)/z_p\sqrt{n}.$$

From the graphs of the non-centrality parameters (τ_L and τ_U) functions shown in Figure 2, it is possible to see that, under H_A when $\mu = 0$ and, consequently, $\sigma=(\Delta-|\mu|)/z_p$, τ_L reaches its minimum value equal to $z_p\sqrt{n}$ and τ_U attains its maximum value of $-z_p\sqrt{n}$.

Focussing our interest on the minimum value, we remember that Pr{E} decreases when τ_L decreases and τ_U increases; therefore, $\inf_{H_A}\Pr\{E\}$ occurs just in the border point $C=(0;z_p\sqrt{n})$, vertex of the ABC triangle of the H_A space, shown in Figure 1 (Panel B).

Therefore, the inferior of the power under H_A is given by:

$$\inf_{H_A}\Pr\{E\}=\inf_{H_A}\Pr\left\{\begin{matrix}(T_L>k_{1-\alpha})\wedge(T_U<-k_{1-\alpha})\\ |\tau_L=z_p\sqrt{n};\tau_U=-z_p\sqrt{n}\end{matrix}\right\}=$$

Formula 8

$$=\Pr\left\{\begin{matrix}(t_{n-1,\tau_L}>k_{1-\alpha})\wedge(t_{n-1,\tau_U}<-k_{1-\alpha})\\ |\tau_L=z_p\sqrt{n};\tau_U=-z_p\sqrt{n}\end{matrix}\right\}$$

A more immediate alternative proof is based on the fact that $\inf_{H_A}\Pr\{E\}$ has to be calculated on the boundary of H_A; therefore, since both conditions must be satisfied, it must be:

$$\theta_{1-p}=-\Delta \text{ and } \theta_p=\Delta.$$

Remembering the definitions of θ_{1-p} and θ_p, this formulation corresponds to a linear system of two equations in the unknowns μ and σ, whose solution is: $\mu = 0$ and $\sigma=\Delta/z_p$.

This conclusion is in agreement with Shieh's affirmation [16] that the power of the test in the point with coordinates $(\mu=0;\sigma=\Delta/z_p)$ is not inferior to a prefixed value 1-β, considering fixed the remaining parameters of the H_A space.

Finally, it has to be noted that $\sup_{H_0}\Pr\{E\}$ corresponds to the maximum significance level (usually 0.05) and that $\inf_{H_A}\Pr\{E\}$ corresponds to the minimum required power value (usually 0.8).

Sample size calculation

It has to be remembered that the null hypothesis (H_0) is a hypothesis of non-equivalence (non-agreement) being the thresholds (±Δ) delimiting the interval considered of equivalence or "of practical equality"; therefore, values greater than $-\Delta_0$ or lower than Δ_0 allow to reject the non-equivalence H_0 hypothesis.

Otherwise, the non-equivalence H_0 hypothesis can be formulated in the terms of a population proportion; accordingly, a population proportion p_0 (or, equivalently, of a central population portion $P_0 = 2\ p_0 - 1$) corresponds to the threshold of the non-equivalence and values greater than p_0 allow to reject the non-equivalence H_0 hypothesis.

It has to be stressed that p_0 or Δ_0 have to be appropriately chosen according to clinical/laboratory considerations based on literature findings.

Furthermore, it is well-known that, in the sample size calculation, a specific alternative hypothesis (a sub-space of H_A) has to be settled in the H_A space delimited by the ABC triangle which characterizes our case of interest (Figure 1, Panel B).

A) Setting the alternative hypothesis according to the two proposed different approaches

Case 1: population proportion p or quantile z_p fixed (different Δ thresholds)

The population proportions are kept fixed ($p_A = p_0 = p$) and lower Δ_A thresholds are selected ($\Delta_A < \Delta_0$), as it is shown in Figure 3 (Panel A).

Accordingly, the alternative hypothesis becomes:

$$H_A:\left(-\Delta_A<\theta_{1-p}\right)\wedge\left(\theta_p<\Delta_A\right).$$

Case 2: thresholds Δ fixed (different population proportions $p_A > p_0$)

The thresholds are kept fixed $\Delta_0=\Delta_A=\Delta$ and a greater p_A population proportion (or quantile z_{pA}) is selected: namely, $p_A > p_0$ (or $z_{pA} > z_{p0}$).

The alternative hypothesis becomes: $H_A:\left(-\Delta<\theta_{1-p_A}\right)\wedge\left(\theta_{p_A}<\Delta\right)$; its space is the hacthed area with vertical lines delimitated by the ABC* triangle shown in Figure 3 (Panel A and Panel B).

It has to be noted that Case 2 corresponds to the Shieh's approach [16, page 5], based on two population proportion values under H_0 and H_A, implicitly considering Δ as fixed.

For both the above considered Cases (Case 1 and Case 2), the sample size is calculated accordingly to the fulfilment of the conditions for controlling the Type I and Type II errors.

B)-Determination of the non-centrality parameters

A fundamental point is the fact that the sample size calculation depends on the calculation of Pr{E} that, in its turn, depends on the non-centrality parameters of the bivariate t distribution.

B.1)-The non-centrality parameters under H_θ

The non-centrality parameters have to be calculated as a limit along a determined direction at the point of the supremum of Pr{E}, that is at the point B = (Δ;0), or, indifferently, at the point A = (-Δ;0) of the ABC triangle, as we have shown in the previous paragraph "Calculation of $\sup_{H_0}\Pr\{E\}$".

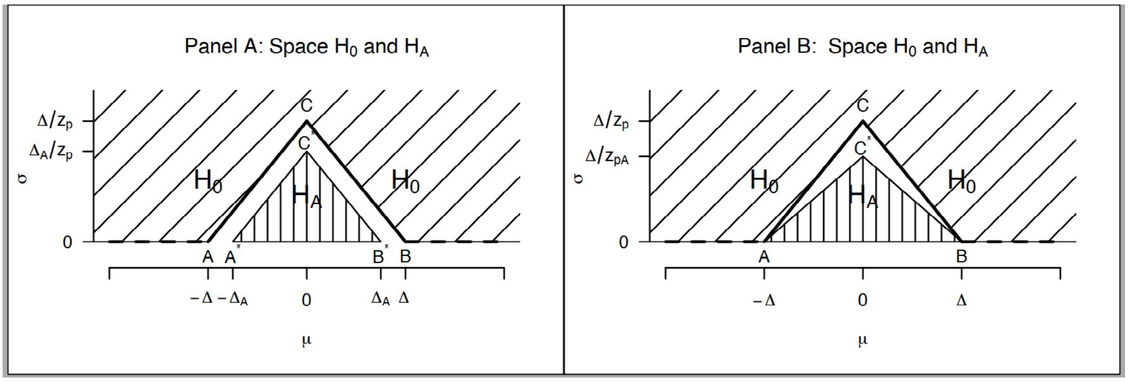
Thus, $\tau_{L_{H_0}}=+\infty$ and $\tau_{U_{H_0}}=-z_p\sqrt{n}$ at the point B or $\tau_{L_{H_0}}=z_p\sqrt{n}$ and $\tau_{U_{H_0}}=-\infty$ at the point A of the ABC triangle.

B.2)-The non-centrality parameters under H_A

The non-centrality parameters have to be calculated at the point of the infimum of Pr{E}, that is, as already reported, at the vertex C* of the space H_A (inner points of the ABC triangle).

Since this vertex is different in the two above

Figure 3. *Panel A: Case 1: population proportion p or quantile z_p fixed (different Δ thresholds); Panel B: Case 2: thresholds Δ fixed (different population proportions $p_A > p_0$)*



outlined cases, the non-centrality parameters will consequently be different.

Case 1: model with fixed population proportion p_0

In this case, $p_0 = p_A = p$ and $C^* = (\mu_{C^*}; \sigma_{C^*}) = (0; \Delta_A / z_p)$.

The non-centrality parameters become:

$$\tau_{L_{H_A}}(C^*) = (\mu_{C^*} + \Delta) / (\sigma_{C^*} / \sqrt{n}) \\ = \Delta / (\Delta_A / z_p)\sqrt{n} = \Delta / \Delta_A \cdot z_p \sqrt{n}.$$

$$\tau_{U_{H_A}}(C^*) = (\mu_{C^*} - \Delta) / (\sigma_{C^*} / \sqrt{n}) \\ = -\Delta / (\Delta_A / z_p)\sqrt{n} = -\Delta / \Delta_A \cdot z_p \sqrt{n}$$

It should be noted that the above non-centrality parameters depend only on p, Δ/Δ_A, and n.

Case 2: model with fixed Δ

In this case, $\Delta_0 = \Delta_A = \Delta$, and $C^* = (\mu_{C^*}; \sigma_{C^*}) = (0; \Delta / z_{p_A})$

The non-centrality parameters become:

$$\tau_{L_{H_A}}(C^*) = (\mu_{C^*} + \Delta) / (\sigma_{C^*} / \sqrt{n}) \\ = \Delta / (\Delta / z_{pA})\sqrt{n} = \Delta / \Delta_A \cdot z_{pA} \sqrt{n}$$

$$\tau_{U_{H_A}}(C^*) = (\mu_{C^*} - \Delta) / (\sigma_{C^*} / \sqrt{n}) \\ = -\Delta / (\Delta / z_{pA})\sqrt{n} = -z_{pA} \sqrt{n}$$

It should be noted that the above non-centrality parameters depend only on p_A and n.

C)-Sample size calculation procedure

The sample size (n) is calculated according to an iterative procedure by using the non-central bivariate t distribution with correlation equal to 1. We have to obtain a "n" value that satisfies the two conditions of $\sup_{H_0} \Pr\{E\} = \alpha$ and $\inf_{H_A} \Pr\{E\} = 1 - \beta$.

Particularly:

a. A starting very low n value (n=2) is iteratively increased until the calculated $k_{1-\alpha}$ gives:

$$\sup_{H_0} \Pr\{E\} = \Pr\left\{ \begin{array}{l} \left(t_{\nu,\tau_{L_{H_0}}} > k_{1-\alpha}\right) \wedge \\ \Pr\left(t_{\nu,\tau_{U_{H_0}}} < -k_{1-\alpha}\right) \end{array} \right\} = \alpha$$

b. Then, we calculate the power:

$$\inf_{H_A} \Pr\{E\} = \Pr\left\{ \begin{array}{l} \left(t_{\nu,\tau_{L_{H_A}}} > k_{1-\alpha}\right) \wedge \\ \Pr\left(t_{\nu,\tau_{U_{H_A}}} < -k_{1-\alpha}\right) \end{array} \right\}$$

c. Furthermore, we increase (or decrease) a low(high) "n" value and we repeat the steps a) and b) until the power reaches the 1- β required threshold with ν = n - 1, $\tau_{L_{H_0}}, \tau_{U_{H_0}}, \tau_{L_{H_A}}$, and $\tau_{U_{H_A}}$ appropriately calculated for the Case 1 or for the Case 2.

Furthermore, it is possible to make more efficient the procedure at the level a) since, as demonstrated in the Appendix B, the $\sup_{H_0} \Pr\{E\}$ can be calculated by using a non-central univariate Student's distribution with ν degrees of freedom and non-centrality parameter $\tau = z_p\sqrt{n}$. Then, our $k_{1-\alpha}$ calculated from the bivariate t distribution corresponds to the 100(1-α)th percentile of the non-central univariate Student's t distribution.

RESULTS

Tables of the sample size

As was demonstrated in the previous paragraph, the sample size calculation, fixed the Type I and Type II errors (α, β), does not depend on μ_0, σ_0, μ_A, and σ_A, but only on p_0 and the ratio Δ_A/Δ_0 in the Case 1 and on p_0 and p_A in the Case 2.

Therefore, the sample size tables have been built by considering two values of the significance level (α = 0.05 or α = 0.01) for two power values (1-β = 0.80 or 0.90). Then, for the Case 1 (Tables 1.1 and 1.2) some selected values of the parameters described above are: p_0 = {0.800, 0.900, 0.950, 0.975, and 0.990 as an extreme value for the non-equivalence} and Δ_A/Δ_0 = {0.80, 0.70, 0.65, 0.60, and 0.55}. Furthermore, for the Case 2 (Tables 2.1 and 2.2), the selected values are: p_0 = {0.800, 0.900, 0.950, 0.975, and 0.990} and p_A = {0.80, 0.90, 0.95, 0.975, and 0.99}.

For example, at a fixed p_0 = 0.9, a null hypothesis of Δ_0 = 1 against Δ_A = 0.8 (Δ_A/Δ_0 = 0.8) is rejected at a significance level of 0.05 and at a power of 0.8 with the specimens obtained from 169 units (Table 1.1, left).

The sample size becomes 214 if the power is increased to 0.90 (Table 1.2, right). Of course, at fixed values of p_0, the sample size decreases at decreasing values of the ratio Δ_A/Δ_0. Finally, the value of Δ_0 = 1 has been considered for simplicity, but any value is possible as long as the value of the ratio is maintained.

For example, H_0: p_0 = 0.9 vs. H_A: p_A = 0.95, is rejected at a significance level of 0.05 and at a power of 0.8 with the specimens obtained from 134 units (Table 2.1, left).

The sample size becomes 169 if the power is increased to 0.90 (Table 2.2, right). Of course, for this approach the values of p_0 and p_A have to be specified.

Table 1.1. Sample size with fixed p (Case1): α = 0.05

Power	0.80					0.90				
Δ_A/Δ_0	0.80	0.70	0.65	0.60	0.55	0.80	0.70	0.65	0.60	0.55
p_0=0.800	282	102	67	46	32	355	127	83	57	40
0.900	169	63	42	30	21	214	79	53	37	27
0.950	134	51	35	25	18	171	65	44	31	22
0.975	118	45	31	22	17	151	58	39	28	21
0.990	106	41	29	21	15	136	53	36	26	19

Table 1.2. Sample size with fixed p (Case1): α = 0.01

Power	0.8					0.9				
Δ_A/Δ_0	0.80	0.70	0.65	0.60	0.55	0.80	0.70	0.65	0.60	0.55
p_0=0.800	215	77	51	35	24	279	100	65	44	31
0.900	128	48	32	22	16	168	62	41	29	21
0.950	101	38	26	19	14	134	50	34	24	17
0.975	88	34	23	17	12	117	45	30	22	16
0.990	79	31	21	16	12	106	41	28	20	15

Table 2.1. Sample size with fixed Δ (Case 2): α = 0.05

Power	0.80					0.90				
p_A	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990
p_0=0.800	*	71	25	15	10	*	88	31	18	12
0.900		*	134	44	22		*	169	55	27
0.950			*	220	54			*	282	69
0.975				*	201				*	259
0.990					*					*

**indicates that the sample size →∞*

Table 2.2. Sample size with fixed Δ (Case 2): α = 0.01

Power	0.80					0.90				
p_A	0.800	0.900	0.950	0.975	0.990	0.800	0.900	0.950	0.975	0.990
p_0=0.800	*	54	19	11	8	*	69	24	14	9
0.900		*	101	33	16		*	132	43	21
0.950			*	166	41			*	220	53
0.975				*	152				*	202
0.990					*					*

**indicates that the sample size →∞*

Comparisons between the sample sizes calculated with p fixed (Case 1) and Δ fixed (Case 2)

The power $\left(\inf_{H_A} \Pr\left\{E\right\}\right)$ increases when τ_L increases and τ_U decreases, keeping all other parameters (α, Δ, and n) fixed. Therefore, to compare the sample sizes calculated for the two considered cases, it is sufficient to compare the corresponding non-centrality parameters, under H_A, indicated shortly as $\tau_{(.)}(1)$ and $\tau_{(.)}(2)$, for the Case 1 and the Case 2, respectively.

The power of the Case 1 is greater or equal to the power of Case 2 when: $\tau_L(1) \geq \tau_L(2)$ and $\tau_U(1) \leq \tau_U(2)$, that is $(\Delta/\Delta_A)*z_p\sqrt{n} \geq z_{pA}\sqrt{n}$, or, equivalently, $\Delta/z_{pA} \geq \Delta_A/z_p$. Consequently, the sample size of Case 1 is less than or at most equal to Case 2.

In geometrical terms, these quantities correspond to the ordinates of the respective points C* under H_A in the Figure 3 (Panel A and Panel B, respectively) equal to $y_{C^*}(1) = \Delta_A / z_p$; $y_{C^*}(2) = \Delta / z_{p_A}$.

Therefore, the sample size of Case 1 is less than or at most equal to the sample size of the Case 2 iff $y_{C^*}(1) \leq y_{C^*}(2)$. In other terms, it is sufficient to compare the two y_{C^*} values: the lower y_{C^*} corresponds to the greater power and the smaller sample size. Finally, the two sample sizes will be equal when $\Delta_A/z_p = \Delta/z_{pA}$.

We show a numerical example with α = 0.05 and power = 0.80.

Case 2 with: p_0 = 0.80 $(z_{p_0} = 0.8416)$, Δ_0 =1, and p_A = 0.90 $(z_{p_A} = 1.2816)$, we have: $\Delta_0 / z_{p_A} = 0.7803$ and $n = 71$;

Case 1 with: p_0 = 0.80 $(z_{p_0} = 0.8416)$, Δ_0 =1, and Δ_A = 0.70, we have: $z_{p_A} = 1.2023 \rightarrow p_A = 0.8854$, $\Delta_0 / z_{p0} = 1.1882$ and n = 102;

Case 1 with: p_0 = 0.80 $(z_{p_0} = 0.8416)$, Δ_0 = 1, and Δ_A = 0.65, we have:

$$z_{p_A} = 1.2023 \rightarrow p_A = 0.8854, \quad \Delta_0 / z_{p0} = 1.1882, and\ n = 67;$$

Then, for obtaining under the Case 1 the same sample size calculated under the Case 2 it is sufficient to determine the value of Δ_A from the equation: $\Delta_A/z_{p0} = \Delta_0/z_{pA}$.

Consequently, we have for the Case 2, p_0 = 0.80 $(z_{p_0} = 0.8416)$, Δ_0 = 1, and p_A = 0.90 $(z_{p_A} = 1.2816)$ giving $\Delta_A = \left(\Delta_0 / z_{pA}\right) \cdot z_{p0} = \left(1/1.2816\right) \cdot 0.8416 = 0.6567$. Therefore, $\Delta_A / z_{p0} = 0.7803$ and n = 71.

A particular approach: the sample size calculation under two simple hypotheses

It has to be pointed out that our new AC procedure considers complex hypotheses on the full H_0 and H_A spaces. However, taking into account that Pr{E}, being fixed the other parameters (α, 1-β, Δ, and n), is function of the pair (μ;σ), it is possible to fix a pair (μ;σ) under H_0 and a pair under H_A and calculate Pr{E} under the two simple hypotheses, together with their corresponding sample size. This sample size calculation, particularly useful for a further verification of the theoretical results previously outlined, differs from the one outlined in the previous paragraph "Sample size calculation procedure" because $\Pr\{E|H_0\}$ and $\Pr\{E|H_A\}$ are now calculated at the fixed pair $(\mu_0;\sigma_0)$ and $(\mu_A;\sigma_A)$ without searching for the supremum and the infimum probability values, respectively.

Consequently, the non-centrality parameters under H_0 are:

$$\tau_{L_{H_0}} = \left(\mu_0 + \Delta / z_{p_0}\right) / \left(\sigma_0 / \sqrt{n}\right) \text{ and}$$

$$\tau_{U_{H_0}} = \left(\mu_0 - \Delta / z_{p_0}\right) / \left(\sigma_0 / \sqrt{n}\right)$$

and, under H_A, are:

$$\tau_{L_{H_A}} = \left(\mu_A + \Delta / z_{p_A}\right) / \left(\sigma_A / \sqrt{n}\right) \text{ and}$$

$$\tau_{U_{H_A}} = \left(\mu_A - \Delta / z_{p_A}\right) / \left(\sigma_A / \sqrt{n}\right)$$

It is possible to calculate these non-centrality parameters directly when are known the points $(\mu_0;\sigma_0)$ and $(\mu_A;\sigma_A)$ and a starting n value. Then, the pertinent sample sizes for comparing two simple hypotheses can be iteratively calculated.

Choosing the points $(\mu_0;\sigma_0)$ and $(\mu_A;\sigma_A)$ appropriately, it is possible to obtain the sample sizes calculated with our new AC procedure and those calculated by Shieh [16] by suitably choosing the two simple hypotheses.

Let's make an example according to our procedure with the following parameters values: α = 0.05, 1 – β = 0.80, Δ= 1, p_0 = 0.90, and p_A = 0.95. If we use our procedure, we obtain a sample size of 134 (Table 2.1). The same sample size of 134 is obtained if we formulate as simple hypothesis H_0 the coordinates of the point A, namely: H_0: $(\mu_0;\sigma_0) = (-\Delta^+;0^+) = \lim_{\mu \to -\Delta^+} \left(\mu; \left(\Delta - |\mu|\right) / z_{p0}\right)$ and as simple hypothesis H_A the coordinates of the point C*, namely: H_A: $(\mu_A;\sigma_A) = \left(0; \Delta / z_{pA}\right)$.

The same result is obtained if we consider the coordinates of the point B, under H_0: $(\mu_0;\sigma_0) = (\Delta^-;0^+) = \lim_{\mu \to -\Delta^-} \left(\mu; \left(\Delta - |\mu|\right) / z_{p0}\right)$.

In addition, if as simple hypotheses H_0 and H_A, we consider the coordinates of the points $C = \left(0; \Delta / z_{p0}\right)$ and $C = \left(0; \Delta / z_{pA}\right)$ respectively, according to Shieh [16], the sample size becomes 62, equal to the value calculated by Shieh's procedure [16].

Simulation studies

These simulation studies have been carried out for having an experimental confirmation of our theoretical conclusions.

A) Check of the fulfilment of the nominal significance level α

It must verify that $\sup_{H_0} \Pr\{E\} = \alpha$.

The fixed parameters of this simulation study are: α = 0.05, Δ = 1, p = 0.90, with p =(P+1)/2, where P is the population central proportion under H_0, that Shieh [16] calls "Null proportion"; otherwise, the parameter μ_0 has been made to vary within the interval -Δ,Δ (precisely, for selected values from -0.99 to 0 and from 0 to 0.99) and, consequently, $\sigma_0 = \left(\Delta - |\mu|\right) / z_p$.

For each of the seven pair of (μ_0;σ_0) under H_0, 10,000 samples of size n = 50 have been simulated and the proportion of H_0 rejection (corresponding to the Type I error and defined by Shieh [16] "simulated proportion") has been calculated according to Shieh's procedure [16] and to our procedure. The results are shown in Table 3.

It is possible to see that for μ_0 = 0 the results are in agreement with Shieh's affirmation [16] that his procedure controls adequately the Type I error. Indeed, this proportion, calculated according to Shieh's procedure [16], is very near to the nominal significance level of α = 0.05, while the Type I error proportion calculated according to our new AC procedure is much lower. However, when differs slightly from 0, the Type I error proportion of Shieh's procedure [16] increases to values much greater than the nominal significance level of 0.05 until a maximum of 0.2222 or 0.2176 (Table 3) for μ_0 values of -0.99 or 0.99, very near to the value of Δ = -1 or Δ = +1 corresponding to the points A or B of the ABC triangle.

Otherwise, the significance level α from our procedure reaches the nominal value and it remains in the required validity intervals [0.04, 0.06] and [0.0450, 0.0550] proposed by Cochran [26] and Bradley [27], respectively.

So, the simulation results confirm, as we expected, the theoretically obtained results. In conclusion, only the sample size calculated according to our procedure allows to respect the nominal significance level α.

B) Check of the fulfilment of the nominal power (1 - β)

It needs to be verified that $\inf_{H_A} \Pr\{E\} = 1 - \beta$.

We limited ourselves in checking the actual test power under H_A only in Case 2 with fixed Δ.

We considered the same scenarios shown by Shieh [16] (Table 5, page 5): that is, nominal power = 0.80 or 0.90, population central proportion (P_0) under H_0 equal to 0.80, 0.90, and 0.95 and, consequently, p_0 = (1+ P_0)/2, central population proportion (P_A) under H_A equal to 0.90, 0.95, and 0.99 and, consequently, p_A = (1+ P_A)/2, Δ = 1, and significance level α = 0.05.

For each scenario, we calculated the pertinent sample size (n) according to our new procedure and we generated, under H_A, 10,000 samples of n units with μ_A = 0 and $\sigma_A = \Delta / z_{p_A}$ (see Case 2).

The proportion of H_0 rejection corresponds to the "simulated power" (Tables 4.1 and 4.2), according to Shieh's terminology [16], and it has to be compared with the "estimated power" calculated from the sample size according to Shieh's procedure [16] and our procedure. Of course, the "estimated power" has to be very near to the required power under which the sample size has been calculated.

For easiness of comparison with the sample sizes, the simulated and estimated power shown in the Shieh's paper [16] have been reported in italic between brackets in the pertinent columns of the Tables 4.1 and 4.2.

It is possible to see that only the sample sizes obtained with our procedure fulfil the expected nominal power.

Table 3. Type I error rates of the agreement test: simulation study

μ_0	σ_0	Simulated proportion Our new procedure - AC	Simulated proportion Shieh procedure
0.00	0.780304146	0.0042	0.0487
0.10 (-0.10)	0.702273731	0.0245 (0.0235)	0.1625 (0.1662)
0.30 (-0.30)	0.546212902	0.0507 (0.0489)	0.2176 (0.2222)
0.50 (-0.50)	0.390152073	0.0507 (0.0489)	0.2176 (0.2222)
0.80 (-0.80)	0.156060829	0.0507 (0.0489)	0.2176 (0.2222)
0.90 (-0.90)	0.078030415	0.0507 (0.0489)	0.2176 (0.2222)
0.99 (-0.99)	0.007803041	0.0507 (0.0489)	0.2176 (0.2222)

Probability of the population centile p = 0.90 obtained from a population central proportion value of P = 0.80, Δ =1, nominal significance level α = 0.05 and µ varying from da 0 to -0.99 (a value near to -Δ) or from 0 to 0.99 (a value near to Δ).

Comparisons among the different considered procedures: our new AC procedure, Shieh [16], Liu and Chow [17], B&A [2], Lin et al. [18] and Lu et al. [10].

All proposals can be formulated in a unifying way by considering the two sample distributions of:

$\bar{D} - k_{1-\alpha} S / \sqrt{n}$ and $\bar{D} + k_{1-\alpha} S / \sqrt{n}$ considered by Shieh [16].

However, it is worthwhile to underline that the values of T_{BL} and T_{BU}, shown in the formulas 22 (B&A's procedure [2]), 24 (Lin et al.'s procedure [18]), and 26 (Lu et al.'s procedure [10]) of Shieh's paper [16], have to be calculated with the quantiles $\hat{\theta}_L$ and $\hat{\theta}_U$, shown after equations 18 and 19 of Shieh's paper [16], instead of the quantiles $\hat{\theta}_{bL}$ and $\hat{\theta}_{bU}$ reported in Shieh's paper [16].

The probability Pr{E} = Pr{"reject H_0"}, on which the sample sizes substantially depend, can be formulated as:

$$\Pr\{E\} = \Pr\left\{ \begin{array}{l} \left(-\Delta < \bar{D} - k_{1-\alpha} S / \sqrt{n}\right) \wedge \\ \left(\bar{D} + k_{1-\alpha} S / \sqrt{n} < \Delta\right) \end{array} \right\}$$

By using the non-central bivariate t distribution:

$\Pr\{E\} = \Pr\{(k_{1-\alpha} < T_L) \wedge (T_U < -k_{1-\alpha})\}$ where

$T_L = (\bar{D} + \Delta) / (S / \sqrt{n})$ and

$T_U = (\bar{D} - \Delta) / (S / \sqrt{n})$

Table 4.1. Calculated sample size, "simulated power", and "estimated power" of the exact agreement test for Δ=1, significance level α 0.05, and power =0 .80

Nominal Power	Null proportion	Alternative proportion	Sample size our new procedure *(SS from Shieh)*	Simulated power *(Shieh)*	Estimated power *(Shieh)*	Difference *(Shieh)*
0.8	0.8	0.90	134 *(62)*	0.8022 *(0.3809)*	0.8028 *(0.3832)*	-0.0006 *(-0.0023)*
0.8	0.8	0.95	44 *(21)*	0.8081 *(0.3998)*	0.8096 *(0.3963)*	-0.0015 *(0.0035)*
0.8	0.8	0.99	16 *(9)*	0.8337 *(0.4791)*	0.8288 *(0.4747)*	0.0049 *(0.0044)*
0.8	0.9	0.95	220 *(118)*	0.8005 *(0.4745)*	0.8008 *(0.4753)*	-0.0003 *(-0.0008)*
0.8	0.9	0.99	32 *(18)*	0,8086 *(0.4857)*	0.8064 *(0.4858)*	0.0022 *(-0.0001)*
0.8	0.95	0.99	78 *(47)*	0.8029 *(0.5456)*	0.8037 *(0.5413)*	-0.0008 *(0.0043)*

Table 4.2. Calculated sample size, "simulated power", and "estimated power" of the exact agreement test for Δ=1, significance level α = 0.05, and power = 0.90

Nominal Power	Null proportion	Alternative proportion	Sample size our new procedure (*SS from Shieh*)	Simulated power (*Shieh*)	Estimated power (*Shieh*)	Difference (*Shieh*)
0.9	0.8	0.90	169 *(86)*	0.9021 *(0.5542)*	0.9006 *(0.5552)*	0.0015 *(-0.0010)*
0.9	0.8	0.95	55 *(28)*	0.9057 *(0.5544)*	0.9050 *(0.5509)*	0.0007 *(0.0035)*
0.9	0.8	0.99	19 *(11)*	0.9089 *(0.6030)*	0.9062 *(0.5999)*	0.0027 *(0.0031)*
0.9	0.9	0.95	282 *(163)*	0.9022 *(0.6467)*	0.9010 *(0.6457)*	0.0012 *(0.0010)*
0.9	0.9	0.99	40 *(24)*	0.9019 *(0.6568)*	0.9019 *(0.6480)*	0.0000 *(0.0088)*
0.9	0.95	0.99	100 *(64)*	0.8998 *(0.6990)*	0.9031 *(0.7048)*	-0.0033 *(-0.0058)*

So, the comparison among the sample sizes for the different procedures becomes the comparison among the different values of the coefficient $k_{1-\alpha}$.

Table 5 shows the values of "k" of the formulas from the different considered procedures.

Table 6 shows the $k_{1-\alpha}$ values for $\alpha = 0.05$ and $p=0.95$ at some sample sizes values. It has to be noted that, within each approach, the values of the sample size and the values of $k_{1-\alpha}$ increase accordingly.

Furthermore, within each sample size value, Shieh's $k_{1-\alpha}$ [16] is the lowest, but, as we have previously theoretically demonstrated and verified by simulation studies, the Type I and the Type II are not fulfilled. Then, in increasing order there are: the $k_{1-\alpha}$ from Lin et al. [18], the $k_{1-\alpha}$ from B&A [2], and the $k_{1-\alpha}$ from our new AC procedure, surprisingly, at first glance, equal the $k_{1-\alpha}$ coefficient obtained from Liu and Chow [17]. Furthermore, the coefficient $k_{1-\alpha}$ from Lu et al. [10] is less than our coefficient at n = 30 and more at n = 200; this pattern also applies to the sample size functions shown in Figure 4 which reports the sample size functions of all the considered approaches.

It is well evident that, among the considered procedures, the sample size function of Shieh's procedure [16] has the lowest values, but it has been calculated under a restrictive H_0, as we have already reported. In addition, our sample size function is between those of Lu et al. [10] more liberal and B&A [5] more conservative. Lin et al. [18] function is practically superimposed on that of B&A [2]. The sample size function of Liu and Chow's procedure [17] does not appear since it is completely superimposed by the sample size function of our procedure owing to the fact that they are equivalent, as we will demonstrate. Finally, it has to be noted that for fixed values of p_0, when the p_A values increase (p_A > 0.98, in Figure 4) the sample sizes decrease, becoming progressively very similar.

Check of the fulfilment of the nominal significance level α of the various methods

Considering Shieh's affirmation [16] about the conservatism of the TOST methods, we have carried out a simulation study to empirically estimate their significance level.

Table 5. Values of "k" of the formulas from the different considered procedures

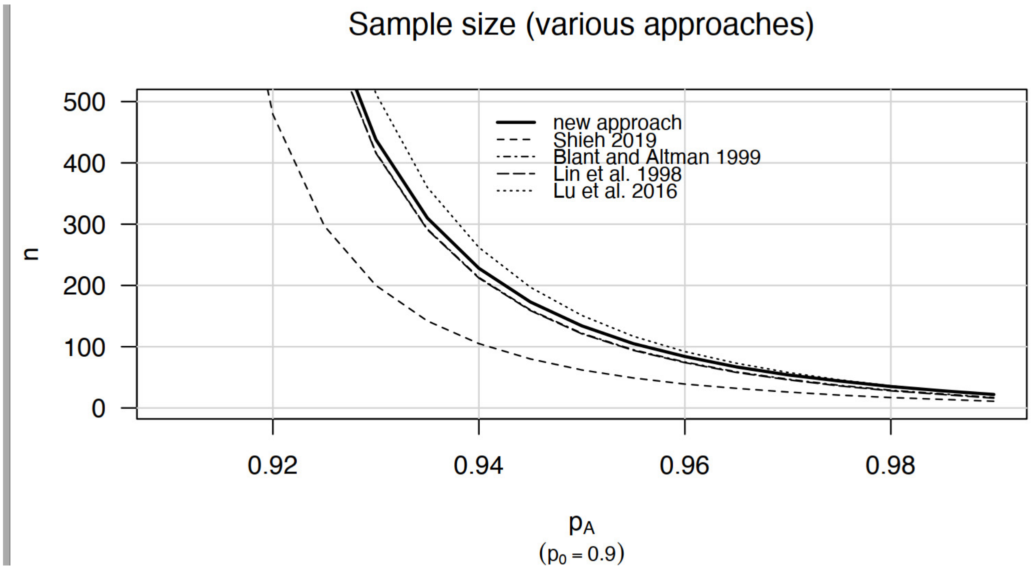
Methods	Coefficient $k_{1-\alpha}$
Our new procedure - AC	Solution $k_{1-\alpha}$ of: $\sup_{H_0} \Pr\left\{E\right\} = \alpha$
Shieh [16]	Solution $\gamma_{1-\alpha}$ of: $Pr\left\{E \mid \mu = 0; \sigma = \Delta / z_p\right\} = \alpha$
Liu and Chow [17]	$t_{1-\alpha,n-1}\, z_p\sqrt{n}$
B&A [5]	$w_{1-\alpha} = z_p\sqrt{n} + t_{1-\alpha,n-1}\sqrt{b}$
Lin et al. [18]	$z_{1-\alpha}\sqrt{b} + z_p\sqrt{n}$
Lu et al. [10]	$t_{1-\alpha/2,n-1}\sqrt{b} + z_p\sqrt{n}$

Where $b = 1 + z_p^2 / 2$

Table 6. $k_{1-\alpha}$ values for $\alpha = 0.05$ and p=0.95 at some sample sizes values

Procedures\ sample size	30	50	100	200
Our new procedure - AC	12.15855	14.60171	19.26539	25.98244
Shieh [16]	10.74401	13.24633	17.96149	24.71098
Liu and Chow [17]	12.15855	14.60171	19.26539	25.98244
B&A [5]	11.61548	14.20249	18.99537	25.79654
Lin et al. [18]	11.53223	14.15387	18.97154	25.78474
Lu et al. [10]	12.14636	14.71331	19.49208	26.28648

Figure 4. Sample Size functions for p_0 =0.90 with p_A > 0.90 of the four considered procedures



We used the same parameter values used by Shieh for obtaining the results shown in the Tables 2, 3 and 4 of his paper [16].

Particularly, we used four sample size values (n = 30, 50, 100, 200), three values of the population proportion (p = 0.900, 0.950,0.975), and, finally Δ = 1. We changed Shieh's [16] μ = 0 and σ = Δ/ z_{p0} with μ = (0.9999 or -0.9999) and σ = (Δ-μ)/ z_{p0}, and, finally, we simulated 100,000 samples. The results are shown in Tables 7.1, 7.2, and 7.3.

It is possible to see that Shieh's procedure [16] is too much liberal with α values ranging from 0.14504 to 0.23524 with an increasing trend, in agreement with the sample size increase. Therefore, Shieh's procedure [16] does not protect from the Type 1 error, as we have already shown (Table 3).

Our procedure and the equivalent Liu and Chow' procedure [17] show the best performance with a values ranging from 0.05012 to 0.05078; interestingly, the higher values (>0.05) are at the two extremes of the sample sizes and the lower values (<0.05) are at the intermediate sample sizes almost giving the impression of a curvilinear trend.

Furthermore, B&A's procedure [2] gives somewhat liberal α values ranging from 0.06131 to 0.08546 with a decreasing trend at the sample size increase.

Finally, Lu et al.'s procedure [10] tends to be somewhat/moderately conservative with α values ranging from 0.03401 to 0.05328 with only two values greater than 0.05 at the lowest sample size (n = 20).

Considerations about the canonical null hypothesis H_0 and Shieh's approch [16]

As reported in the Methods paragraph, H_0 is the union of two hypotheses: H_{0A} H_{0B}.

Let's consider now a null hypothesis (H_0^*) more restrictive, corresponding to the intersection of the two hypotheses H_{0A} and H_{0B} (H_{0A} H_{0B})

$H_0^* : \theta_{1-p} \leq -\Delta$ *and* $\theta_p \geq \Delta$ in which "*and*" has replaced the "or" of the canonical formulation of H_0.

It has to be noted that H_0^* is contained in the null canonical hypothesis (H_0) as it is possible to see in Figure 5. The space of H_0^* corresponds to the region delimited by the extensions of the sides AC and BC of the triangle ABC in the half plane positive of the vertical Y axis (Y > 0) and it is well evident that is a subset of the space outside the ABC triangle of the canonical H_0 formulation.

According to this formulation, the individual agreement test should be formulated as:

$H_0^* : \theta_{1-p} \leq -\Delta$ and $\theta_p \geq \Delta$ vs. $H_A : -\Delta < \theta_{1-p}$ and $\theta_p < \Delta$

To determine the sample size according to H_0^*, it is necessary to carry out the same steps outlined for our new procedure in the canonical case, as previously reported.

a)-Determination of the supremum of Pr{E} under H_0.

Similarly to the canonical case, the $\sup_{H_0} \Pr\{E\}$ has to be calculated on the boundary of H_0 that is, on the extensions of the sides AC and BC in the direction of the positive Y, but, unlike the canonical case, both conditions must be satisfied, i.e. it must be both: $\theta_{1-p} = -\Delta$ and $\theta_p = \Delta$.

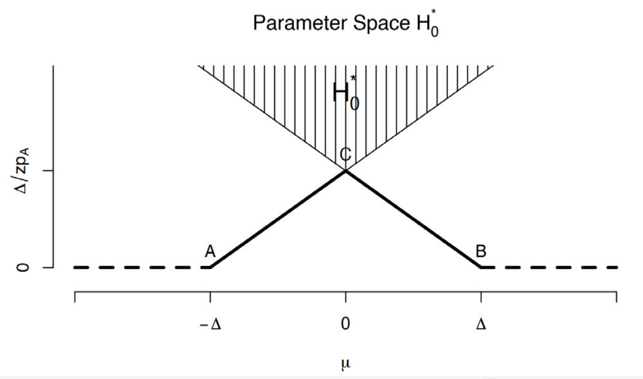
Remembering the definitions of θ_{1-p} and θ_p, this corresponds to a linear system of two equations in the unknowns μ and σ, whose solutions are μ = 0 and σ = Δ/z_p the same ones reported by Shieh [16], corresponding to the vertex C of the triangle ABC.

Figure 5. Parameter space of the H_0 hypothesis*

Table 7.1	n=30	n=50	n=100	n=200
Our new procedure - AC	0.05031	0.04954	0.04938	0.05012
Shieh [16]	0.21658	0.22286	0.22895	0.23524
Liu and Chow [17]	0.05031	0.04954	0.04938	0.05012
B&A [5]	0.07958	0.07118	0.06526	0.06131
Lin et al. [18]	0.08596	0.07476	0.06708	0.06206
Lu et al. [10]	0.04791	0.04210	0.03647	0.03401

Table 7.2	n =30	n=50	n=100	n=200
Our new procedure - AC	0.05068	0.04978	0.04927	0.05031
Shieh [16]	0.17106	0.17548	0.18220	0.18723
Liu and Chow [17]	0.05068	0.04978	0.04927	0.05031
B&A [5]	0.08312	0.07379	0.06676	0.06255
Lin et al. [18]	0.08949	0.07755	0.06852	0.06356
Lu et al. [10]	0.05123	0.04436	0.03787	0.03447

Table 7.3	n=30	n=50	n=100	n=200
Our new procedure - AC	0.05078	0.04980	0.04941	0.05057
Shieh [16]	0.14505	0.14879	0.15511	0.15926
Liu and Chow [17]	0.05078	0.04980	0.04941	0.05057
B&A [5]	0.08546	0.07559	0.06794	0.06390
Lin et al. [18]	0.09171	0.07929	0.06972	0.06484
Lu et al. [10]	0.05328	0.04562	0.03860	0.03538



b)-Determination of the infimum of Pr{E} under H_A

For this calculation, the H_A hypothesis is the same as previously reported and, consequently, the $\inf_{H_A} \Pr\{E\}$ is on the point with coordinates μ = 0; $\sigma = \Delta/z_p$ corresponding to the vertex C of the triangle ABC, similarly as it has been shown by Shieh [16].

Then, if we modify our procedure by using these values in the calculation of $\sup_{H_0} \Pr\{E\}$ and $\inf_{H_A} \Pr\{E\}$ or we use the procedure for the sample size calculation under two simple hypotheses (see paragraph titled "A particular approach: the sample size calculation under two simple hypotheses."), we obtain the same sample sizes as Shieh [16].

In conclusion Shieh's procedure [16] is valid only in the case of the more restrictive hypothesis H_0*, which is a part of the canonical null hypothesis H_0. We have to restate that our procedure allows to obtain the sample size pertinent to the general case.

Considerations about Liu and Chow's TOST procedure [17] and our AC procedure

The explanation of the fact that these two procedures give the same sample size is based on the calculation of the coefficient $k_{1-\alpha}$ and of the test power. Indeed, our procedure calculates the coefficient $k_{1-\alpha}$ by referring to the non-central bivariate t distribution, while Liu and Chow [17] use a non-central univariate Student's t distribution. However, we demonstrate in the Appendix B that the calculation of the coefficient $k_{1-\alpha}$ using the non-central bivariate t distribution is equivalent to using the non-central univariate Student's t distribution.

Otherwise, regarding the power calculation, both procedures make use of the non-central bivariate t distribution. Thus, being equal the two quantities on which the sample size calculation is based, it is expected that the sample size values will be the same, fixed the parameters of the sample size calculation.

A relevant difference that has to be noted, is the fact that our procedure is based on a test of size α i.e. with a Type I error equal to α. Liu and Chow [17] claimed that "the proposed two one-sided tests procedure for individual bioequivalence is size α test, but using the notation for "only a level α test (i.e. ≤α)" according to Casella and Berger's definition [20]. However, for completeness, we will demonstrate in the Appendix C that Liu and Chow's test [17] is just a size α test (i.e. =α), according to Casella and Berger's definition [20].

DISCUSSION

The sample size calculation for agreement studies has to be appropriately done starting from sensible assumptions in addition to the usual power and significance level (two tailed).

Indeed, we think that all the studies need to be supported by a sample size calculation justified by a sound statistical methodology in order to have an adequate probability of obtaining their aim. So, the sentence: "All studies need a sample size justification. Not all studies need a sample size calculation." from the Han et al.'s review [1] is, in our opinion, difficult to be justified and shareable. Indeed, we think that researchers have to be well aware of the effect sizes, differences, etc. that even a sample size of convenience allows to demonstrate.

Furthermore, we think that an approach based on the individual equivalence owing to its sensible rationale for assessing the agreement between measurement methods has to be absolutely supported even if the estimates of their systematic and proportional biases have to be absolutely supplied. As a further relevant point, sample sizes have to be such as to not compromise the real feasibility of an agreement study taking into account also the ethic aspects of the biomedical research [28].

A relevant point to be noted is that the sample sizes calculated according to Shieh [16] or our procedure does not depend on the values of μ and σ under the H_0 and H_A hypotheses.

Indeed, the sample size calculated with our procedure, fixed α, 1 – β and p_0, depends in the Case 1 only on the ratio Δ*/Δ and, in the Case 2, only on p_A,whatever the value assigned to Δ.

This rather surprising result occurs also in Shieh's sample size calculation [16], fixed α, 1 – β, p_0 and p_A; indeed, the sample size is always the same independently from the value given to Δ, as it is possible to calculate from the IML® or R programs attached to Shieh's paper [16].

Moreover, the sample size from the proposed our procedure is obtained with an exact statistical method similarly to the sample size obtained by Shieh's procedure [16], instead of the approximate methods followed by B&A [2], Lin et al. [18], and Lu et al. [10].

In addition, our procedure allows to assess the individual bioequivalence and to calculate its pertinent sample size according to two different, but complementary approaches that are easy to switch between. Indeed, the first uses different Δ thresholds (Δ_0, Δ_A) and the second uses different quantiles (z_{p0}, z_{pA}) or their corresponding probabilities p (p_0, p_A), under H_0 and under H_A. As a further relevant point, the sample sizes calculated for sensible agreement scenarios are adequate for the actual feasibility of the study.

As another relevant point, our approach considers the whole parameter space of H_0 and H_A hypotheses and allows to obtain a test of size α according to Casella and Berger [20].

Our procedure gives the same sample sizes of the Liu and Chow's procedure [17], leading to conclude that the two procedures are equivalent and also that Liu and Chow's procedure [17] is of size α as we have also directly demonstrated. Finally, it has to be stressed that these procedures have a better performance in comparison to the other considered procedures owing to their better control of the Type I and Type II errors, as we have theoretically demonstrated and as our simulation study has empirically confirmed (Tables 7.1, 7.2. and 7.3).

So, our procedure or Liu and Chow's procedure [17] has to be warmly recommended.

We have demonstrated that Shieh's proposal [16] is based on a particular case of the parameters space under H_0 and, consequently, it has some very important limitations. Particularly, under the canonical formulation of the H_0 and H_A hypotheses, it does not give the claimed statistical significance test of size α and do not fulfil the required power, being its sample sizes too much lower than the necessary ones.

However, it is debatable whether it is possible to shrink the H_0 space to Shieh's formulation [16] without fulfilling the axiom of the complementarity between the null and the alternative hypotheses.

REFERENCES

1. Bland M, Altman A (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476): 307-310.
2. Bland MJ, Altman D (1999) Measuring Agreement in Method Comparison Studies. Stat Methods Med Res 8(2):135-160.
3. Han O, Tan HW, Julious S, Sutton L, Jacques R, et al. (2022) Review of samples sizes used in agreement studies published in the PubMed repository BMC Med Res Methodol 22(1): 242.
4. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, et al. (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64(1): 96-106.
5. Gerke O, Pedersen AP, Debrabant B, Halekoh U, Möller S (2022) Sample size determination in method comparison and observer variability studies. J Clin Monit Comput 36(5): 1241-1243.
6. Bland JM, Altman D (1990) A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement

Comput Biol Med 20(5): 337-340.
7. Taffè P When can the Bland & Altman limits of agreement method be used and when it should not be used. J Clin Epidemiol 2021;137:176-181.
8. Bland JM Website https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm
9. Cesana BM, Antonelli P (2012) Agreement analysis: further statistical insights. Ophthal Physl Opt 32(5): 436-440.
10. Lu MJ , Zhong WH, Liu YX, Miao HZ, Li YC, et al. (2016) Sample Size for Assessing Agreement Between Two Methods of Measurement by Bland-Altman Method. The International Journal of Biostatistics 12(2): 1-8.
11. Shieh G (2016) Exact Power and Sample Size Calculations for the Two One-Sided Tests of Equivalence Plos One 11(9): e0162093.
12. Shieh G (2018) The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. BMC Med Res Methodol 18(1): 45.
13. Carkeet A (2015) Exact Parametric Confidence Intervals for Bland-Altman Limits of Agreement. Optom Vis Sci 92(3): e71-80.
14. Carkeet A, Goh YT (2018) Confidence and coverage for Bland-Altman limits of agreement and their approximate confidence intervals. Stat Methods Med Res 27(5):1559-74.
15. Jan SL, Shieh G (2018) The Bland-Altman range of agreement: Exact interval procedure and sample size determination. Comput Biol Med 100:247-252.
16. Shieh G (2020) Assessing Agreement Between Two Methods of Quantitative Measurements: Exact Test Procedure and Sample Size Calculation. Stat Biopharm Res 12(3):352-359.
17. Liu JP, Chow SC (1997) A Two One-Sided Tests Procedure for Assessment of Individual Bioequivalence. J Biopharm Stat 7(1): 49-61.
18. Lin SC, Whipple DM, Ho CS (1998) Evaluation of statistical equivalence using limits of agreement and associated sample size calculation. Commun Stat A-Theor 1998;27(6): 1419-1432.
19. Krishnamoorthy K, Mathew T Statistical Tolerance Regions: Theory, Applications, and Computation. 2009, New York: Wiley. (page 35, equation 2.3.11).
20. Casella G, Berger RL Statistical Inference, 1990, Duxbury Press, Belmont 2nd edition, 2002 Duxbury Pacific Grove Ca USA
21. Owen DB (1965) A special case of a bivariate non-central t-distribution. Biometrika 52(3-4): 437-446.
22. Dunnett CW, Sobel MA (1954) Bivariate Generalization of Student's t-distribution, with Tables for certain special cases. Biometrika 41(1-2): 153-169.
23. https://CRAN.R-project.org/package=OwenQ
24. https://CRAN.R-project.org/package=PowerTOST
25. Genz A, Bretz F, Miwa T, Mi X, Leisch F, et al. (2021) mvtnorm: Multivariate Normal and t Distributions. R package version 1.1-3.
26. Cochran WG (1954) Some Methods for Strengthening the Common $\chi 2$ Tests. Biometrics 10(4): 417-451.
27. Bradley JV (1978) Robustness? Brit J Math Stat Psy 31(2): 144-152.
28. Cesana BM, Antonelli P (2019) Sample size calculations need to be adequate and parsimonious. J Clin Epidemiol 108:140-141.

Appendix A: demonstration that T_L and T_U are non-central Student's t distributions.

First of all, let's remember that: $T_L = \left(\bar{D}+\Delta\right)/\left(S/\sqrt{n}\right)$ and $T_U = \left(\bar{D}-\Delta\right)/\left(S/\sqrt{n}\right)$.

Considering the value of T_L, it is possible to write:

$$T_L = \frac{\left(\bar{D}+\Delta\right)}{S/\sqrt{n}} = \frac{\left(\bar{D}+\mu-\mu+\Delta\right)}{S/\sqrt{n}} \cdot \frac{\sigma/\sqrt{n}}{\sigma/\sqrt{n}} = \left[\frac{\left(\bar{D}-\mu\right)}{\sigma/\sqrt{n}} + \frac{\left(\mu+\Delta\right)}{\sigma/\sqrt{n}}\right] \cdot \frac{\sigma/\sqrt{n}}{S/\sqrt{n}}$$

Where μ and σ refer to the distribution of the differences (D).

Putting $\tau_L = \frac{\left(\mu+\Delta\right)}{\sigma/\sqrt{n}}$ and noting that $\frac{\left(\bar{D}-\mu\right)}{\sigma/\sqrt{n}} = Z$ (the standardized Gaussian distribution) and $\frac{(n-1)S^2}{\sigma^2} = \chi^2_{n-1}$ (a χ^2 distribution with n-1 degrees of freedom independent from Z), we obtain:

$$T_L = \frac{Z+\tau_L}{S/\sigma} = \frac{Z+\tau_L}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{Z+\tau_L}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

That is, by definition a non-central univariate Student's t distribution with ν = n - 1 degrees of freedom and non-centrality parameter given by τ_L, formally: $T_L \sim t_{\nu,\,\tau L}$.

Similarly, considering the value of T_U, we obtain:

$$T_U = \frac{Z+\tau_U}{S/\sigma} = \frac{Z+\tau_U}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \text{ with } \tau_U = \frac{\mu-\Delta}{\sigma/\sqrt{n}}$$

Formally, $T_U \sim t_{\nu,\,\tau U}$.

Appendix B

Demonstration that the calculation of sup$_{H0}$ P{E} can be done by means of a non-central univariate Student's t distribution and determination of the k$_{1-\alpha}$ coefficient corresponding to its 100(1-α)th percentile.

Owen's Q functions

Non-central univariate Student's t

Owen21 introduced the following Q-functions:

$$Q_1(t,\nu,\delta,R):=\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_0^R \Phi\left(\frac{tx}{\sqrt{\nu}}-\delta\right)x^{\nu-1}e^{-x^2/2}dx$$

$$Q_2(t,\nu,\delta,R):=\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_R^{+\infty} \Phi\left(\frac{tx}{\sqrt{\nu}}-\delta\right)x^{\nu-1}e^{-x^2/2}dx$$

where R is every value >0.

Furthermore, the cumulative probability function of a non-central univariate Student's t with ν degrees of freedom and non-centrality parameter δ, is defined as:

$$F(t,\nu,\delta):=\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_0^{+\infty} \Phi\left(\frac{tx}{\sqrt{\nu}}-\delta\right)x^{\nu-1}e^{-x^2/2}dx$$

This function can be split in the sum of two terms, corresponding to the above reported Owen's function, as:

$$F(t,\nu,\delta):=\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_0^R \Phi\left(\frac{tx}{\sqrt{\nu}}-\delta\right)x^{\nu-1}e^{-x^2/2}dx+\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_R^{+\infty} \Phi\left(\frac{tx}{\sqrt{\nu}}-\delta\right)x^{\nu-1}e^{-x^2/2}dx=$$

$= Q_1(t,\nu,\delta,R)+Q_2(t,\nu,\delta,R)$ for each $R > 0$

It follows immediately that, for $R\to+\infty, F(t,\nu,\delta)=Q_1(t,\nu,\delta,+\infty)$

Non-central bivariate t

The cumulative probability function of a non-central bivariate t with ν degrees of freedom for t$_1$>t$_2$ and δ$_1$>δ$_2$, with correlation equal to 1, according to Owen [21] is:

$$\Pr\left(T_1\le t_1 \wedge T_2\le t_2\right)=\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_0^R \Phi\left(\frac{t_1x}{\sqrt{\nu}}-\delta_1\right)x^{\nu-1}e^{-x^2/2}dx+$$

$$+\frac{1}{\Gamma(\nu/2)\cdot 2^{(\nu-2)/2}}\int_R^{+\infty} \Phi\left(\frac{t_2x}{\sqrt{\nu}}-\delta_2\right)x^{\nu-1}e^{-x^2/2}dx$$

with $R=\frac{\delta_1-\delta_2}{t_1-t_2}\sqrt{\nu}$.

So, by using the Owen's Q functions [21]:

$\Pr\left(T_1\le t_1 \wedge T_2\le t_2\right)= Q_1(t_1,\nu,\delta_1,R)+Q_2(t_2,\nu,\delta_2,R)$

From this distribution it is possible to calculate the probabilities of the following events:

$$E_2=\left\{T_1>t_1\wedge T_2\le t_2\right\}, E_3=\left\{T_1\le t_1\wedge T_2\ge t_2\right\}, \text{and } E_4=\left\{T_1\ge t_1\wedge T_2\ge t_2\right\}$$

Particularly, the probability of the event E$_2$, corresponding to our event E, is:

$$\begin{aligned}\Pr\{E_2\} &= \Pr\{T_2 \le t_2\} - \Pr\{T_1 \le t_1 \wedge T_2 \le t_2\} = \\ &= Q_1(t_2,\nu,\delta_2,R) + Q_2(t_2,\nu,\delta_2,R) - [Q_1(t_1,\nu,\delta_1,R) + Q_2(t_2,\nu,\delta_2,R)] = \\ &= Q_1(t_2,\nu,\delta_2,R) - Q_1(t_1,\nu,\delta_1,R)\end{aligned}$$

Then, the probability of the event E, crucial for the sample size calculation, is given by:
$\Pr\{E\} = \Pr\{T_L > k_{1-\alpha} \wedge T_U < -k_{1-\alpha}\} = \Pr\left\{t_{\nu,\tau_L} > k_{1-\alpha} \wedge t_{\nu,\tau_U} < -k_{1-\alpha}\right\}$
where $[t_{\nu,\tau L}; t_{\nu,\tau U}]$ is a non-central bivariate t-distribution with degrees of freedom ν, non-centrality parameters τ_L and τ_U, and with correlation equal to 1.
Then, putting: $t_1 = k_{1-\alpha}, \delta_1 = \tau_L, t_2 = -k_{1-\alpha}, \delta_2 = \tau_U$, for keeping Owen's terminology [21], Pr{E} can be also expressed as: $\Pr\{E\} = Q_1(-k_{1-\alpha},\nu,\tau_U,R) - Q_1(k_{1-\alpha},\nu,\tau_L,R)$

Calculation of $\sup_{H_0}\Pr\{E\}$

Let's remember that the supremum of Pr{E} under H_0 is obtained in the point $(\mu,\sigma) = \left[-\Delta^+; (\Delta - |\mu|)/z_p = 0^+\right]$ leading, consequently, to the following non-centrality parameters: $\tau_L = z_p\sqrt{n}$ and $\tau_U = -\infty$, and, finally, R=+∞.

Otherwise, it can be also obtained in the point $(\mu,\sigma) = \left[\Delta^-; (\Delta - |\mu|)/z_p = 0^+\right]$ with non-centrality parameters given by: $\tau_L = +\infty$ and $\tau_U = -z_p\sqrt{n}$.

So: $\sup_{H_0}\Pr\{E\} = Q_1(-k_{1-\alpha},\nu,-\infty,+\infty) - Q_1\left(k_{1-\alpha},\nu,z_p\sqrt{n},+\infty\right)$

Let's analyse the two terms at the right of the above equality; we can see that:

a) $Q_1(-k_{1-\alpha},\nu,-\infty,+\infty) = 1$.

In fact, in the formula of Q_1 the term $\Phi\left(\frac{tx}{\sqrt{\nu}} - \delta\right)$ becomes Φ(+∞) = 1; consequently, the integrand function of Q_1 becomes a χ density function that, integrating over all its domain, returns a value of 1.

b) From the relationships among Owen's Q functions [21] and the non-central univariate Student's t distribution, previously reported, we can write:

$$Q_1\left(k_{1-\alpha},\nu,z_p\sqrt{n},+\infty\right) = Q_1\left(k_{1-\alpha},\nu,z_p\sqrt{n},R\right) + Q_2\left(k_{1-\alpha},\nu,z_p\sqrt{n},R\right) = F_t\left(k_{1-\alpha},\nu,z_p\sqrt{n}\right)$$

Where $F_t(k_{1-\alpha},\nu,\tau)$ indicates the cumulative distribution of the non-central t with ν degrees of freedom and non-centrality parameter τ.
In conclusion: $\sup_{H_0}\Pr\{E\} = 1 - F_t\left(k_{1-\alpha},\nu,z_p\sqrt{n}\right)$

Calculation of the coefficient $k_{1-\alpha}$

The coefficient $k_{1-\alpha}$ is calculated so that: $\sup_{H_0}\Pr\{E\} = \alpha$. Therefore, it is sufficient to solve the equation $1 - F_t\left(t,\nu,z_p\sqrt{n}\right) = \alpha$ in respect to t, that is $F_t\left(t,\nu,z_p\sqrt{n}\right) = 1 - \alpha$; the solution is exactly the 100(1-α)% centile of a non-central univariate Student's t with ν degrees of freedom and non-centrality parameter $\tau = z_p\sqrt{n}$ given by:

$$k_{1-\alpha} = t_{1-\alpha,\nu,z_p\sqrt{n}}$$

It can be seen that the coefficient $k_{1-\alpha}$ corresponds to the coefficient $\tau_{1-\alpha}$ of the Liu and Chow's procedure [17], described by Shieh [16] (formula 13 at page 3).

Numerical check of the previous theoretical results

In the open-source R language with $p_0 = 0.9$, $p_A = 0.95$, $\alpha = 0.05$, power = 0.80; we obtain, according to our procedure, a sample size of n = 134 with $k_{1-\alpha} = 17.25068$.

1) Verify that $Q_1\left(-k_{1-\alpha},\nu,-\infty,+\infty\right)=1$

```
> OwenQ::OwenQ1(nu = 133, t = -17.25068, delta = -Inf, R = .Machine$double.xmax)
[1] 1
```

It has to be noted that the parameters of OwenQ are: "nu" = degrees of freedom, t = our $k_{1-\alpha}$, delta = our non-centrality parameter, and R a constant in the Owen's formula. In addition, ".Machine$double.xmax" represents the maximum numerical value obtainable in the open-source R language and approximates +Inf.

2) Verify that Owen's Q_1 function is equal to the cumulative probability of the non-central t distribution:

$$Q_1\left(k_{1-\alpha},\nu,z_p\sqrt{n},+\infty\right)=F_t\left(k_{1-\alpha},\nu,z_p\sqrt{n}\right)$$

2.1)- Calculation of $Q_1\left(k_{1-\alpha},\nu,z_p\sqrt{n},+\infty\right)$

```
>OwenQ::OwenQ1(nu= 133,t = 17.25068, delta = qnorm(p = 0.90)*sqrt(134), R = .Machine$double.xmax)
   [1] 0.9499998
```

2.2)- Calculation of $F_t\left(k_{1-\alpha},\nu,z_p\sqrt{n}\right)$ the cumulative probability of a non-central t distribution, using "pt" function of R.

```
> pt(q = 17.25068,df = 133, ncp = qnorm(p = .90)*sqrt(134))
   [1] 0.9499998
```

3) Verify that $k_{1-\alpha}$ is equal to the 1-α centile of the non-central t distribution: $k_{1-\alpha} = t_{1-\alpha,\nu,z_p\sqrt{n}}$

3.1)-Calculation of $k_{1-\alpha}$ by using an ad hoc written function for the new proposed our procedure

```
   > k_coeff(n = 134,alpha = .05,p = .9)["k_coef"]
   [1] 17.25068
```

3.2)-Calculation of $t_{1-\alpha,\nu,z_p\sqrt{n}}$ by using the "qt" function of R

```
   > qt(p = .95, df = 133, ncp = qnorm(p = 0.90)*sqrt(134)) #
   [1] 17.25068
```

Appendix C

Demonstration that the generalized procedure two-one sided test (TOST) for the assessment of the individual agreement is a size α test, in contrast to Shieh's affirmation [16].

In order for it to be a size α test, it is necessary to verify that the two conditions, indicated in theorem 8.3.24 of Casella and Berger's book [20, page 396], are satisfied.

The two tests that make up the TOST procedure are:

$H_{01}:\theta_{1-p}\le-\Delta$ vs. $H_{A1}:\theta_{1-p}>-\Delta$

$H_{02}:\theta_{1-p}\ge\Delta$ vs. $H_{A2}:\theta_{1-p}<\Delta$

Let's us consider the first test that make up the TOST procedure.

Casella and Berger's first condition [20] is that a sequence of parameter points ($\mu_i;\sigma_i$) in H_{01} exists such that:

$$\lim_{i\to+\infty}\Pr\left(\text{reject } H_{01}\,|\left(\mu_i;\sigma_i\right)\right)=\alpha\,.$$

The test function is: $\hat{\theta}_{1-p}=\bar{D}-k\cdot S/\sqrt{n}$, where k is an appropriate constant to be determined.

The rejection Region of H_{01} is $\hat{\theta}_{1-p}\le-\Delta$ and it is equivalent, after some algebraic steps, to: $\left(\bar{D}+\Delta\right)/\left(S/\sqrt{n}\right)>k_{1-\alpha}$.

Given, $T_L=\left(\bar{D}+\Delta\right)/\left(S/\sqrt{n}\right)$, we have demonstrated, in Appendix A, that T_L is distributed as a non-central univariate Student's t with non-centrality parameter $\tau_L=\left(\mu+\Delta\right)/\left(\sigma/\sqrt{n}\right)$.

In order that the partial test is of size α, k must satisfy the following condition:

$$\sup_{H_{01}}\Pr\left\{\hat{\theta}_{1-p}>-\Delta\right\}=\alpha \text{ or, equivalently}: \sup_{H_{01}}\Pr\left\{T_L>k\right\}=\alpha\,.$$

The superior has to be searched on the boundary of H_{01} and, in this particular case, is on the half straight line in the positive half plane of the Y axis on which lies the segment AC.

On the half line AC, $\sigma=\left(\left(\Delta+\mu\right)/z_p\right)$ and, consequently, $\tau_L=\left(\mu+\Delta\right)/\left(\left(\Delta+\mu\right)/\left(z_p\sqrt{n}\right)\right)=z_p\sqrt{n}$

Thus, the condition $\sup_{H_{01}}\Pr\left\{T_L>k\right\}=\alpha$ becomes $\Pr\left\{T_L>k\,|\,\tau_L=z_p\sqrt{n}\right\}=\alpha$.

Since T_L is a non-central Student's t, the coefficient k which fulfils the above condition corresponds to the (1-α)th quantile of this non-central univariate Student's t distribution with non-centrality parameter $\tau_L=z_p\sqrt{n}$ and consequently it will be denoted as $k_{1-\alpha}$.

Thus, the test function becomes $\hat{\theta}_{1-p}=\bar{D}-k_{1-\alpha}S/\sqrt{n}$.

Let's consider a succession of points on the segment AC (see Figure 1, Panel A) belonging to the border of H_{01} (and also of H_0), which from C converge towards A.

We have:

$$\lim_{\mu\to-\Delta^+}\left(\mu;\sigma=\frac{\Delta+\mu}{z_p}\right)=\left(-\Delta^+;0^+\right) \text{ converging to the point A}$$

$$\lim_{\mu\to-\Delta^+}\tau_L=\lim_{\mu\to-\Delta^+}\frac{\mu+\Delta}{\Delta+\mu}z_p\sqrt{n}=z_p\sqrt{n}$$

Thus, $\lim\limits_{\mu \to -\Delta^{+}} \Pr\{T_L > k_{1-\alpha}\} = \Pr\left\{t_{(n-1;\tau_L)} > k_{1-\alpha} \mid \tau_L = z_p\sqrt{n}\right\} = \alpha$.

The last equality is a consequent of what we have just reported about $k_{1-\alpha}$.
Therefore, Casella and Berger's first condition [20] is satisfied.

Casella and Berger's second condition [20] is that a sequence of parameter points $(\mu_i;\sigma_i)$ in H_{01} exists such that: $\lim\limits_{i \to +\infty} \Pr\left(\text{reject } H_{02} \mid (\mu_i;\sigma_i)\right) = 1$.

Let us now consider the second test of the TOST procedure.

$H_{02}: \theta_p \geq \Delta$ vs. $H_{A2}: \theta_p < \Delta$

The test function is: $\hat{\theta}_{1-p} = \bar{D} + k_{1-\alpha} \cdot S / \sqrt{n}$ and the rejection Region of H_{02} is $\hat{\theta}_p < \Delta$ and, given $T_U = \left(\bar{D} - \Delta\right) \big/ \left(S / \sqrt{n}\right)$ it is equivalent to $T_U < -k_{1-\alpha}$, where T_U is distributed as a non-central univariate Student's t with non-centrality parameter $\tau_U = \left(\mu - \Delta\right) \big/ \left(\sigma / \sqrt{n}\right)$.

Let's consider the same succession of points on the segment AC (see Figure 1, Panel A) as we did previously.
We have:

$$\lim_{\mu \to -\Delta^{+}} \tau_U = \lim_{\mu \to -\Delta^{+}} \frac{\mu - \Delta}{\Delta + \mu} z_p \sqrt{n} = -\infty$$

Thus, $\lim\limits_{\mu \to -\Delta^{+}} \Pr\{T_U < -k_{1-\alpha}\} = \Pr\left\{t_{n-1;\tau_U} < -k_{1-\alpha} \mid \tau_U = -\infty\right\} = 1$

The last equality is a consequence of the fact that the non-central univariate cumulative probability of the Student's t distribution is a decreasing function with respect to its non-centrality parameter.
Thus, Casella and Berger's second condition [20] is also satisfied.

We can therefore conclude that Liu and Chow's TOST procedure [17] is a size α test.

Robust Regression as a Sensible Alternative to the Weighted Ordinary Least Squares Regression in case of Heteroskedasticity. A Tutorial

Annalisa Orenti[(1)], *Anna Zolin*[(1)], *Ettore Marubini*[(*)], *Paolo Antonelli*[(2,#)], *Federico Ambrogi*[(1)], *Bruno Mario Cesana*[(1,#)]

(1) University of Milan, Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Biometry and Epidemiology "G.A. Maccacaro", Milan, Italy
(2) Retired Professor of Calculus of Probabilities, Statistics and Operative Research at the State Industrial Technical Institute (ITIS) Benedetto Castelli, Brescia, Italy
(*) deceased
(#) retired

CORRESPONDING AUTHOR: Annalisa Orenti, Department of Clinical Sciences and Community Health, Laboratory of Medical Statistics, Biometry and Epidemiology "Maccacaro", University of Milan. Address: Via Celoria 22, 20133 Milan, Italy. E-mail: annalisa.orenti@unimi.it

SUMMARY

Background: The robust regression is rarely used in the statistical analyses in comparison with the Ordinary Least Squares regression and the Weighted Regression. In addition, in the frequent case of the heteroskedasticity of the residuals, a weighted regression carried out once is the main suggestion of the statistical books and the resulting reduced heteroscedasticity is usually considered sufficiently satisfactory.
Methods: We showed the OLS regression analysis on data simulated with a well evident heteroskedasticity and an ad hoc outlier, followed by a weighted regression iteratively carried out by using iteratively reweighted least squares, an estimation method used also in several procedures of the robust regression analysis. Therefore, the link between the iteratively performed weighted regression and the robust regression becomes immediate. Furthermore, the same data have been analysed using some robust regression procedures.
Results: It has been shown that in a simulated sample of heteroscedastic data with and without an obvious artificially created outlier the weighted regression performs worse with more biased parameter estimates than robust regression procedures (such as the robust MO procedure) as the presence of the outlier is not adequately neutralized.
Discussion: In presence of a heteroskedastic pattern of the residuals, the suggestion to use robust regression procedures which can also deal with the almost sure presence of outliers seems more sensible. Among the robust regression procedures carried out, the performance of the robust MO procedure appears particularly appealing since it allows biostatisticians a more reasoned management of the outliers shown in a very illustrative "ad hoc" plot. Robust regression procedures represent a sensible alternative to OLS regression taking into account that its assumptions are practically not always fulfilled and that outliers, which are almost certainly present, are not only difficult to handle in classical OLS regression but can also provide highly biased estimates.

Keywords: weighted regression; heteroscedasticity; iterative reweighted least squares, robust regression procedures (MO).

DOI: 10.54103/2282-0930/26484
Accepted: 23rd September 2024

INTRODUCTION

Robust regression does not appear adequately used by professional and not professional biostatisticians despite the theoretical advances on robust regression analysis done in the last thirty years as witnessed by excellent comprehensive textbooks from, among others, Atkinson and Riani[1], Rousseeuw and Leroy [2] Maronna et al. [3,4], Huber [5], and Huber and Ronchetti [6].

The aim of this tutorial is to show that, in case of a heteroskedastic pattern of the residuals, robust regression methods can replace much more effectively the weighted regression based on the Iterative ReWeighted Least Squares (IRWLS). In addition, robust regression methods allow optimal management of potential or real outliers.

This proposal can be considered an advance compared to the usual recommendation to resort to a single weighted regression usually suggested by many authors [7,8,9] to remove or at least reduce heteroskedasticity.

We therefore hope that biostatisticians, both professional and non-professional, will be more willing to adopt robust regression techniques also because the commercial and non-commercial software packages available today offer adequate data processing capabilities, capable of managing the computational requirements of these robust procedures.

Readers are assumed to be familiar with the statistical methodology of Ordinary Least Squares Regression (OLS-R) with a focus of methods for testing its assumptions and detecting outliers. In addition, at least a basic knowledge of Weighted Least Squares Regression (WLS-R, considered in more detail in this paper) is required. The relationships between the main equations used by OLS-R and WLS-R are shown in Table S4 of the supplementary material (s.m., thereafter).

Linear Regression - Statistical Theory

In this tutorial we will consider the second type of linear regression with both independent and dependent random variables. In fact, the first or "classic" type with the independent variable as a fixed variable is rather a theoretical model, useful however in the calibration of a new measurement method and the third type with both variables with a measurement error in addition to the biological variability belongs to the so-called measurement error models not considered in this paper. Furthermore, we will consider a simple linear regression (without loss of generality, assuming the number of the independent variables k equal to 1), so that the data points can be displayed in a scatter plot of X and Y to easily investigate the homogeneous or non-homogeneous pattern of the distribution of their sample values. However, our example is easily extendable to multiple regression.

For the Ordinary Least Squares (OLS) method, used to obtain the vector **b**, estimate of the parameter vector **β**, readers can refer to standard regression books such as Draper and Smith [7], Kuther et al. [8] or Chatterjee and Hadi [9] and the section "Linear regression – Statistical Theory" of the s.m. Here we would like to emphasize that the vector **b**, as given by equation (4) in Table S4 of the s.m., is the Best Linear Unbiased Estimator (BLUE) of the regression parameters (intercept and regression coefficients). This property follows from the Gauss-Markov theorem (see Kutner et al. [8, Chapter 1, page 18]) which states that among the unbiased linear estimators of **b**, the estimator with minimum variance is obtained by the Minimum Weighted Squares (MWS) method with the weight matrix (**W**) equal to the inverse of the error variance-covariance matrix: $\mathbf{W}=\mathbf{\Sigma}^{-1}$. Thus, OLS method is a special case of "Weighted Least Squares" (WLS) method when, due to homoscedasticity, equal residual variances can be collected to a common factor leaving all weights equal to 1. In turn, "Weighted Least Squares" method is a particular case of "Generalized Least Squares" (GLS) one when the error variance-covariance matrix is diagonal (i.e., the error terms are uncorrelated) with heteroskedasticity (the weights are different: $w_{i,i}= 1/\sigma_i^2$, being $w_{i,i}$ a generic element on the main diagonal of the weights matrix **W**; in fact, the inverse of a diagonal matrix is equal to a diagonal matrix with the reciprocal of its values on the diagonal).

Furthermore, the unbiasedness of the estimators does not require that the errors be normally distributed nor that they be independent and identically distributed as long as they are uncorrelated with zero mean and homoskedastic with finite variance. However, the requirement of the unbiasedness property has to be maintained since there also exist estimators with lower variance but biased.

Moreover, Carroll and Ruppert [12] do not consider the WLS but only the GLS that have the advantage of being applied without any distributional assumptions but specifying only the model for mean, variances and their relationship.

The OLS estimates are obtained by minimizing the sum of the squared residuals; namely,

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(Y_i-\alpha-\beta X_i\right)^2 \qquad (1)$$

Otherwise, in the case of the WLS the function to be minimized becomes:

$$\sum_{i=1}^{n}w_{i,i}\varepsilon_i^2=\sum_{i=1}^{n}w_{i,i}\left(Y_i-\alpha-\beta X_i\right)^2 \qquad (2)$$

where $w_{i,i}$ is the i-th diagonal element of the (nxn) matrix **W**; the other terms are been defined in the paragraph "Linear Regression – Statistical Theory" of the s.m.. Thus, a weighted sum of the squared residuals is minimized, where each squared residual is weighted by the reciprocal of its variance. In other words, when estimating **b**, less weight is given to the observations for which the linear relationship (to be estimated) is noisier and more weight to those for which it is less noisy.

Unfortunately, after an OLS-R, the recommended steps to test the statistical assumptions of the model (errors: identically, independently and normally distributed – the latter particularly relevant for the validity of the statistical tests on the estimates and the ANOVA table of the regression analysis) and also the adequacy of the model (straight line) are not systematically carried out even by professional biostatisticians.

Therefore, one may not observe a fan-shaped (or megaphone-like) pattern shown by the (externally studentized) residuals plotted against the fitted (predicted) values as an expression of a heteroskedastic distribution (higher variance for higher values of the fitted values) rather than the homoskedastic one required for the validity of the OLS analysis.

Methods: statistical analyses

We performed a simple OLS regression on the simulated dataset according to a heteroskedastic regression model as described in the s.m.. In particular, the mean and the standard deviation of the independent variable (X) were equal to 14 (μ_X) and to 6 (σ_X), respectively. The regression slope (β_1) and intercept (β_0) parameters are both 0.9.

For showing the advantage of performing a robust regression (see after), observation n.4 of the dataset was created as an outlier by increasing its ordinate to 15.09 from the original value of 2.33 and keeping the original simulated abscissa value of 1.59. Figure 1 shows the diagram plot of the simulated data (Panel A) and the diagram plot of the same data with observation n. 4 modified as an outlier (Panel B). In Figure 1, points whose OLS residuals were found outside some thresholds for outlier diagnostics are shown in red and marked with the observation number.

In order not to lose the thread of this tutorial from OLS heteroscedasticity to robust regression passing via weighted regression, the OLS regression results (ANOVA table, parameter estimates together with their standard error, t-statistics with their p-value, mean error squares, coefficient of determination (R^2) not adjusted and adjusted) are shown in the s.m. (Table S1 and Table S1.1, respectively). The paragraph "Considerations about the coefficient of determination" which deals with some theoretical aspects of the unadjusted and adjusted coefficient of determination of OLS and WLS regressions, the paragraphs "outlier diagnostics: theory" and "outlier diagnostics: data with the outlier", together with some plots obtained with the keyword "influence" from the OLS regression by SAS® Proc REG [26] are shown and commented in the s.m. to which interested readers are referred.

Figure 2 shows the plot of the "externally studentized residuals" or "jackknifed residuals" (see s.m.) called "Rstudent" in SAS®, more effective in detecting outlying Y observations than "internally studentized residuals", vs. the fitted values as a practical example of residual heteroskedasticity of the fifty simulated data without and with the artificially created outlier. The points are marked with the observation number to better understand why these observations exceeded some thresholds to be considered "suspected outliers" (see s.m.).

In Figure 2-Panel A, a fan-shaped pattern is quite evident, especially considering that these residuals

Figure 1. Scatter diagram of the simulated data (Panel A) and the same data with observation n.4 modified as an outlier (Panel B)

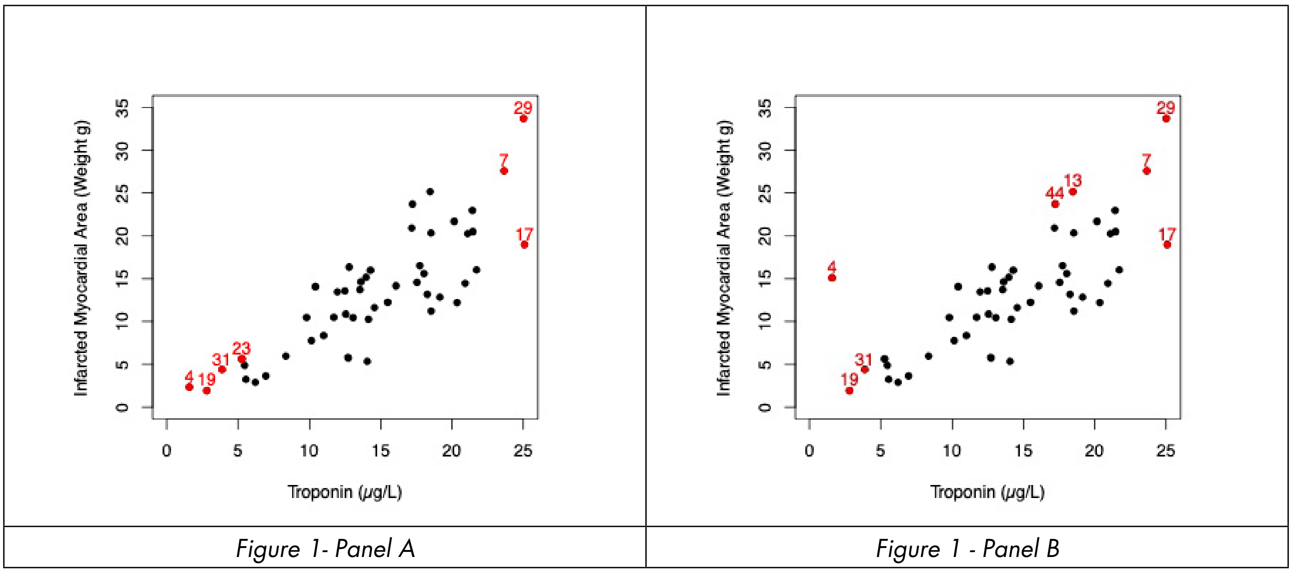


Figure 1- Panel A

Figure 1 - Panel B

Figure 2. Scatter diagram of the externally studentized residuals and fitted values for the simulated data (Panel A) and the same data with observation n.4 modified as an outlier (Panel B)

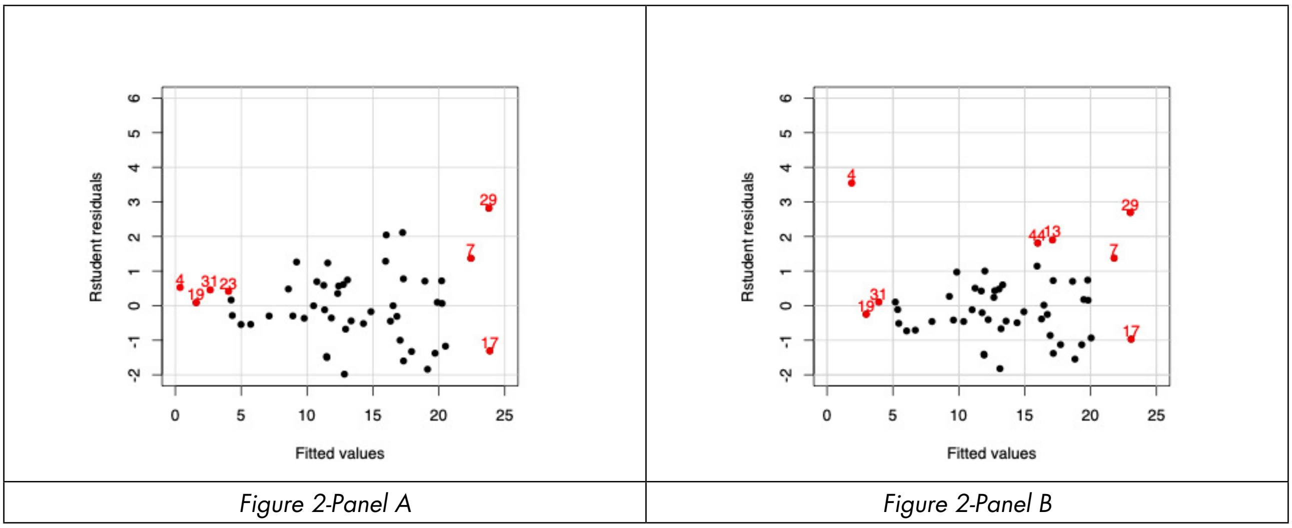


Figure 2-Panel A *Figure 2-Panel B*

come from simulated data with heteroskedastic errors. The same pattern is shown in the Figure 2-Panel B for the data with the artificially created Y-direction outlier, although the outlier (n.4 in the top left of the Figure 2-Panel B) is well above the fan-shaped pattern better evident in Figure 2-Panel A without the outlier.

It should be noted that, due to heteroscedasticity, OLS estimator still provides unbiased and consistent estimates but no longer of minimum variance (see Kutner et al. [8 Chapter 11]). Finally, according to Chatterjee and Hadi [9, 5th Ed. Page 193] it is possible to conclude that the coefficients "lack precision in a theoretical sense".

The presence of the heteroskedasticity can be formally tested by means of some formal statistical tests, as reported in Appendix A of the s.m.. However, since the statistical tests may provide inconsistent results, it is strongly recommended to base the judgement of heteroskedasticity on the pattern of the (externally studentized) residuals against the independent or the dependent or the fitted variables.

Outlier Diagnostics

Another relevant point in any type of statistical analysis, but particularly relevant in the context of the regression analysis, is the identification of outliers (usually defined as Outlier Diagnostics; see s.m.) in the sample dataset. In fact, the presence of just one outlier can dramatically modify the OLS estimates. Furthermore, it should be emphasized that the presence of a heteroskedastic pattern makes more difficult the identification of the outliers. For the outlier diagnostics (leverage, Mahalanobis distance or better the squared Mahalanobis distance -MDi^2-, the "standardized residuals", the "studentized residuals" or the "internally studentized residual", the DFFITS(i) statistics, the DFBETASj(i) statistics, and the "COVRATIO") together with the related diagnostic thresholds, readers are referred to the already cited paragraphs "Outlier diagnostics: theory" and "Outlier diagnostics: data with the outlier" of the s.m.

Statistical approaches to deal with the heteroskedasticity

Heteroskedasticity can be removed by a suitable transformation of the variables according to Cohen et al. [32] and Mosteller and Tukey [33] with their Bulging Rule (or Ladder of Powers) suggesting power transformations (of X or of Y or both) with exponents of 2, 1, 0.5, -0.5, -1 and -2 including the logarithmic transformation. However, this approach leads to the very relevant problem of attributing a meaning to the relationship between the transformed variables with respect to what the researcher wanted to evaluate between the original variables.

However, even non-professional biostatisticians are well aware that, in the case of heteroskedasticity, several statistical books dedicated to regression [7,8,9] suggest the Weighted Regression (WR) analysis as a more sensible alternative to transformations. WR is a procedure based on a generalization of the regression model, which is implemented by assigning different weights to each observation, instead of the weight equal to 1 given by the OLS method, being homoscedasticity. Therefore, it is necessary to calculate weighted least squares (WLS) estimates instead of OLS. Furthermore, the use of weights will (legitimately) impact the widths of the statistical intervals.

However, there is a practical difficulty in determining the weights to estimate the error variances (or standard deviations) that provide the **W** matrix to be used to obtain the WLS estimator ($\mathbf{b}_w$).

In some cases, weights values may be based on theory or previous research. For example, when the error variance is proportional to an independent variable, the natural weights are the reciprocal of the independent variable.

However, in the usual case where the structure of the matrix **W** is unknown, it is necessary to estimate the variance or the standard deviation function, accordingly.

In experiments designed with large numbers of replicates or in the case of some measurements of the dependent variable at the same or nearly the same value of the independent variable (replicates or nearly replicates), weights can be estimated directly from the sample variances of the response variable at each combination of the same or nearly the same values of the independent variable. An exemplification of this approach is shown in Draper and Smith's book [7, paragraph 9.2. Generalized Least Squares and Weighted Least Squares, page 221] where the variances of the Y values, computed at five means of equal or nearly equal X values, were regressed on the X means and the X means squared, due to a suggested quadratic relationship. Of course, this regression allows the variance pattern to be modelled and the regression coefficients to be used obtain the fitted variance values for each X. Then, the reciprocal of the fitted variance at each X value was used as the weight for the WR between the original Y and X dataset.

As a further example of this approach, consider the analysis of the dataset (https://online.stat.psu.edu/stat501/lesson/13/13.1) consisting of seven observations of the pea diameter (in inches) of the parent plant (X) and the mean diameter (in inches) of up to 10 plants grown from seeds of the parent plant (Y) made by Sir Francis Galton (16 February 1822 Birmingham England – 17 January 1911 Haslemere, Surrey, England). Therefore, it is possible to calculate the variance of the progeny plants and use its reciprocal as weights for the WR analysis.

Further issues are reported in Appendix B of the s.m. to which the interested reader is referred.

Usually, statistical books report an example of WR together with the plot of the WLS residuals against the variable for which a megaphone pattern has been highlighted and conclude that a satisfactory or rather satisfactory reduction in heteroskedasticity has been achieved. For example, Draper and Smith [7, page 229] report "The residuals plots in Figure 9.2 reveal that the vertical spread of residuals is now *roughly (our italic)* the same at the two main levels of the transformed response. At lower levels there are only two observations so that there is not much of an estimate of the spread there. The employment of weighted least squares here appears to be justified and useful".

WLS estimates of coefficients are generally close to the "ordinary" unweighted OLS estimates. However, Kutner et al. [8, 1974, page 426] and Chatterjee et al. [9] point out that if the estimated WLS coefficients differ substantially from the estimated OLS coefficients it is recommended to repeat the WLS regression until the estimated coefficients stabilize by using the revised weights obtained by the residual of the previous WLS regression to re-estimate the variance or the standard deviation function; this process gives the "*iteratively reweighted least squares (IRLS or IRWLS)*." Often the stabilization of the coefficients is achieved in no more than one or two iterations.

Similarly, the same procedure should be followed in the case of an unsatisfactory removal of the heteroskedasticity after the first WLS regression, as shown by the plots of the residuals towards the variable with which the megaphone pattern is highlighted at the first OLS-R. Only the studentized residuals take into account the weights that are used to model the different values of the variance and, consequently, these residuals must be used to draw the diagnostic plots.

However, the iterative steps of the weighed regression are a demanding procedure especially if they have to be performed without the availability of ad hoc software.

Indeed, after the externally (better) studentized OLS residuals plotted against a predictor (or fitted values) show a megaphone shape, a second OLS should be performed to estimate the variance function (or the standard deviation function) by regressing the squared residuals (or the absolute residuals) on the predictor or fitted values with which a megaphone pattern was evidenced. Indeed, if the first OLS-R model is correct the i-th squared residual is an estimate of σ_i^2 and the i-th absolute residual is an estimate of σ_i to be preferred in presence of outliers with expected largest residuals so as not to have a very low weight being $w_i=1/\sigma_i^2$ or $1/\sigma_i$.

Then the reciprocal of the fitted values by the estimated variance ($1/\sigma_i^2$) or standard deviation function ($1/\sigma_i$) are used to obtain the weights for the first WLS-R. Next, the procedure needs to be repeated using the set of externally studentized residuals from the WLS-R to re-model the variance (or standard deviation) with an OLS-R and the resulting residuals will be the weights to be used in the second WLS-R, and so on.

An "ad hoc" software that allows to perform the iteratively reweighted least squares procedure could be very useful and in fact a code in the open-source R language is available upon request to the corresponding author.

Table S2 s.m. shows the results of the iterative process using IRLS starting from the OLS regression with a heteroskedasticity pattern clearly evident from the plot of the OLS externally studentized residuals and the fitted Ŷ (Y-hat) variable (Figure S1-Panel A equal to Figure 1-Panel A).

Finally, the OLS regression results are reported in s.m. together with those of the iterative steps of the weighted regression (Figure S1, Panel A, Panel B, Panel C, and Panel D).

Iteratively reweighted Least Squares (IRWLS)

The IRWLS estimator used above in iterative weighted regression is also used in some "robust regression" procedures, among other estimating approaches.

However, in weighted regression the parameter estimates are obtained by minimizing the weighted sum of the squared residuals; otherwise in the robust regression the parameter estimates are obtained by minimizing a particular function of the squared residuals.

Robust regression instead of the OLS regression allows to dampen the influence of outliers that inevitably exist in medical and biological datasets and that are difficult for the researcher to handle.

In fact, it is worth remembering the following clarification of Maronna et al. [3, page 51]: "… while in the classical approach to statistics one aims at estimates which have desirable properties at an exactly specified model, the aim of robust methods is loosely speaking to develop estimates which have a 'good' behaviour in a 'neighbourhood' of a model".

An online SAS® documentation reports the connection between robust regression and weighted least squares. [34] In fact, the use of the IRWLS estimator naturally leads to considering a robust regression whose main advantage consists in an adequate handling of outliers.

Other relevant references are the papers of Holland and Welsch [35], Street, Carroll and Ruppert [36] Heiberger and Becker [37], and Green [38].

Readers interested in a more in-depth clarification of the robust regression methodology and its estimating procedure may referred to Appendix B of s.m..

Since the weights change from one iteration to another, the *weighted* residual sum of squares could not decrease at each iteration. Indeed, for removing this restriction, it has to specify the keyword "NOHALVE" in the PROC NLIN of SAS® [26].

We would like to consider as an explanatory approach the robust Multiple Options (MO) procedure proposed by Orenti and Marubini [39] (see Appendix C of s.m. for an in-depth illustration) together with the robust regression results obtained from the Least Trimmed Squares (LTS) and the MM estimator.

The MO procedure has the particular characterization of iteratively checking for outliers after obtaining a first bulk of observations and has demonstrated satisfactorily behaviour for identifying outliers. The others two robust procedures were considered for their satisfactorily performance and frequent use.

Table 1. Results on data with the outlier

Estimates	OLS	OLS*	IRWLS	MO	MM	LTS
Intercept (b_0) SE P	-1.22785 1.49385 0.4152	0.42970 1.67672 0.7988	1.6778 1.4586 0.256	-0.65877 1.45311 0.652	-0.6064 1.6354 0.712	0.16334 1.47560 0.912
Slope (b_1) Se P	1.00084 0.09645 <.0001	0.90325 0.10825 <0.0001	0.8133 0.1056 <0.0001	0.95421 0.09482 <0.0001	0.9520 0.1303 <0.0001	0.85023 0.09634 <0.0001
MSE	3.94245	4.42505	4.495	3.535	4.072	3.576
R^2	0.6917	0.5919	0.5528	0.6784	0.6303	0.6338
Adj-R^2	0.6853	0.5834	0.5435	0.6717	0.6226	0.6257

OLS - OLS regression with the observation n.4 made as an outlier owing to the increase of the simulated value of its ordinate.*

OLS, WLS, and Robust regression procedure results

Table 1 shows the results of the fitted regression models. The estimates of the weighted regression performed with the iterative weighted estimators and of the MO, MM, and LST robust regressions are those computed at the final step of their iterative process.

Column 2 (OLS) and column 3 (OLS*) allow to capture the large difference between the OLS intercept estimates without and with the observation n. 4 created as an outlier. Indeed, it was expected that the outlier in the Y-direction located approximately at the beginning of the observations would shift the regression line clockwise leading to a positive intercept rather than a negative one. Furthermore, since there is only one outlier the OLS slope estimate is relatively little affected with a reduction of about 10%.

The IRWLS estimates of the weighted regression are very influenced by the presence of only one outlier with the greatest positive intercept and the lowest slope values leading us to conclude that, at least in this case, the weighted regression with IRWL was not a sensible choice.

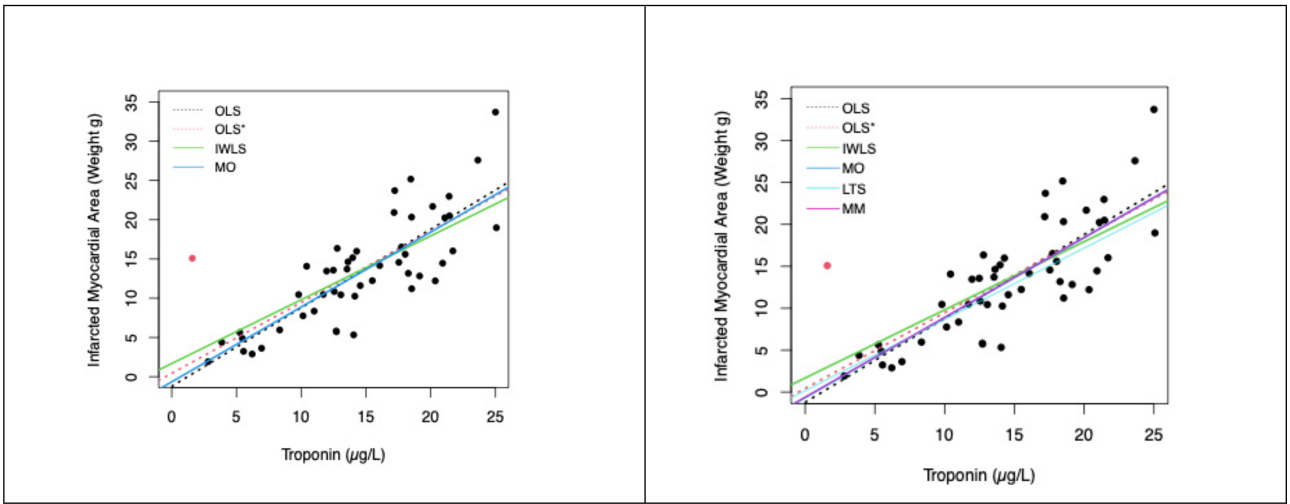


Figure 3-Panel A
OLS = OLS regression with the outlier*
OLS = OLS Regression without the outlier
(observation n.4 with original data)

Figure 3-Panel B. The regression lines of the LTS and MM estimator of the robust regression are in addition shown

Furthermore, the estimates of the MO procedure are very close to those of the OLS without the outlier; finally, taking as a reference the OLS estimates of the dataset without the outlier, the MO estimates turned out to be a little less biased than the MM and LTS estimates.

Figure 3-Panel A shows the regression lines fitted with the OLS, the OLS*, IRWLS, and MO robust regression. Figure 3-Panel B shows also the regression lines fitted from the LST and MM robust regressions. This figure makes it easier to understand the comments reported about the Table 1 with the OLS* regression line deviating by the created outlier and the MO, LTS and MM regression lines close to the OLS line. Finally, it is evident that the weighted regression with the IRWLS estimator (line with the greatest intercept) is not able to overcome the influence of the outlier.

Table 2 shows the results of the regression models fitted without the outlier. The first (OLS) and the second (OLS*) column are equal to the corresponding columns in Table 1 and have been reported for easier comparison. The estimates of the weighted regression performed with the iterative weighted estimators and of the MO, MM, and LST robust regressions are those calculated in the final step of their iterative process.

Again, it is worth highlighting the poor performance of the IRWLS weighted regression with the and the lower bias of the robust MO procedure compared to the other two robust procedures.

Figure 4 shows the regression lines fitted with the OLS, IRWLS, MO, LTS, and MM robust regressions. In particularly, due to their very similar intercept and slope estimates the MO and MM regression lines overlap: specifically, the intercept is -0.78189 and -0.7723, and the slope is 0.96160 and 0.9607, respectively.

Table 2. Results from data without the outlier

Estimates	OLS	OLS*	IRWLS	MO	MM	LTS
Intercept (b_0) SE P	-1.22785 1.49385 0.4152	0.42970 1.67672 0.7988	0.19188 0.49302 0.699	-0.78189 1.33895 0.562	-0.7723 1.2122 0.527	0.2768 1.2859 0.831
Slope (b_1) SE P	1.00084 0.09645 <.0001	0.90325 0.10825 <0.0001	0.88750 0.05756 <0.0001	0.96160 0.08818 <0.0001	0.9607 0.1128 <0.0001	0.8433 0.0851 <0.0001
MSE	3.94245	4.42505	2.285	3.435	3.859	3.325
R^2	0.6917	0.5919	0.832	0.7124	0.6627	0.6858
Adj-R^2	0.6853	0.5834	0.8285	0.7064	0.6659	0.6788

OLS - OLS regression with the observation n.4 created as an outlier by increasing the simulated value of its ordinate*

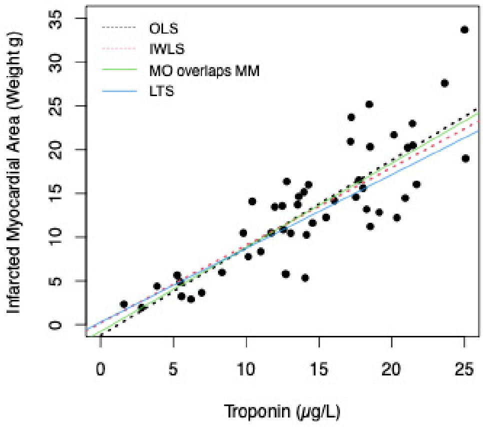


Figure 4

Figure 5, Panel A and Panel B shows the very informative plot obtained from the final iteration of the robust MO regression [39]. Indeed, Figure 5 shows four quadrants: two quadrants are above the horizontal line drawn at 0.5 which delimits the value of weight given to the observations (under or above) and the two remaining quadrants are to the right and left of the vertical line drawn at 0.89517, a value given by the Neperian logarithm of the square root of the 0.95 quantile of a χ^2 distribution with 2 degrees of freedom (ln√5.99146 = ln (2.44775) = 0.89517) (see Appendix C of the s.m.). This line corresponds to the threshold of the Neperian logarithm of the robust distance (ln $_Z$RD) that delimits the low (left) and high (right) leverage points. In particular, in the two quadrants below the line drawn at the weight value of 0.5 there are the observations considered as outliers and "bad leverage points" for the observations on the right of the vertical line drawn at the above reported value of 0.89517. In addition, the two quadrants over the horizontal line drawn at the weight value of 0.5 are the location of the "bulk" (left quadrant) and of the "good leverage points" (right quadrant) as opposed to the "bad leverage points" as they influence the regression fitting without providing biased estimates compared to those that would be obtained with the bulk data.

It is possible to see two observations in the dataset with the outlier (n.4 and n. 29 with weights of 0.16241 and 0.45452, respectively) under the line of the 0.5 threshold in the bottom right quadrant. Of course, in the dataset without observation n.4 created as an outlier, only observation n. 29 is considered an "outlier" with an attributed weight of 0.45155. According to MO, observation n. 7 is at a relevant distance (exceeding the above reported threshold of Cook's robust distance) from the bulk for both datasets as a "good leverage point".

Moreover, observations n. 13, and 44 are close to or above the threshold for the dataset with the outlier and show a further shift to the right for the dataset without the outlier. For the latter dataset, observation n.13 becomes a "good leverage point" while observation n.4 and observation n. 44 are close to but to the left of the threshold and just at the threshold, respectively.

The robust MO procedure assigns weights to the observations without the imputed outlier with a mean of 0.9110 (±0.1139, s.d.), median equal to 0.9674, first (Q_1) and third (Q_3) quartiles equal to 0.8647 and 0.9894, respectively, minimum and maximum values of 0.4516 (obs. n.29) and 1.0000, respectively. Only two observations (n. 19 and n. 28) have a weight of 1, contrary to what happens in the OLS regression in which all observations have a weight of 1. In addition, even in this case the negative skewness of the distribution is evident since the median is much greater than the arithmetic mean.

The weights given by MO to the observations with the imputed outlier have mean of 0.9010 (±0.1548, s.d.), median equal to 0.9654, first (Q_1) and third (Q_3) quartiles equal to 0.8681 and 0.9894, respectively, minimum and maximum values of 0.1624 (obs. n. 4) and 0.9999 (obs. n. 28), respectively. No observations have a weight of 1, contrary to what happens in the dataset without the observation n.4 created as an outlier and in the OLS regression. In addition, the negative skewness of the distribution is evident since the median is much greater than the arithmetic mean.

The descriptive statistics of the weights of the other three methods (IRWLS, MM, and LTS) are reported in the s.m..

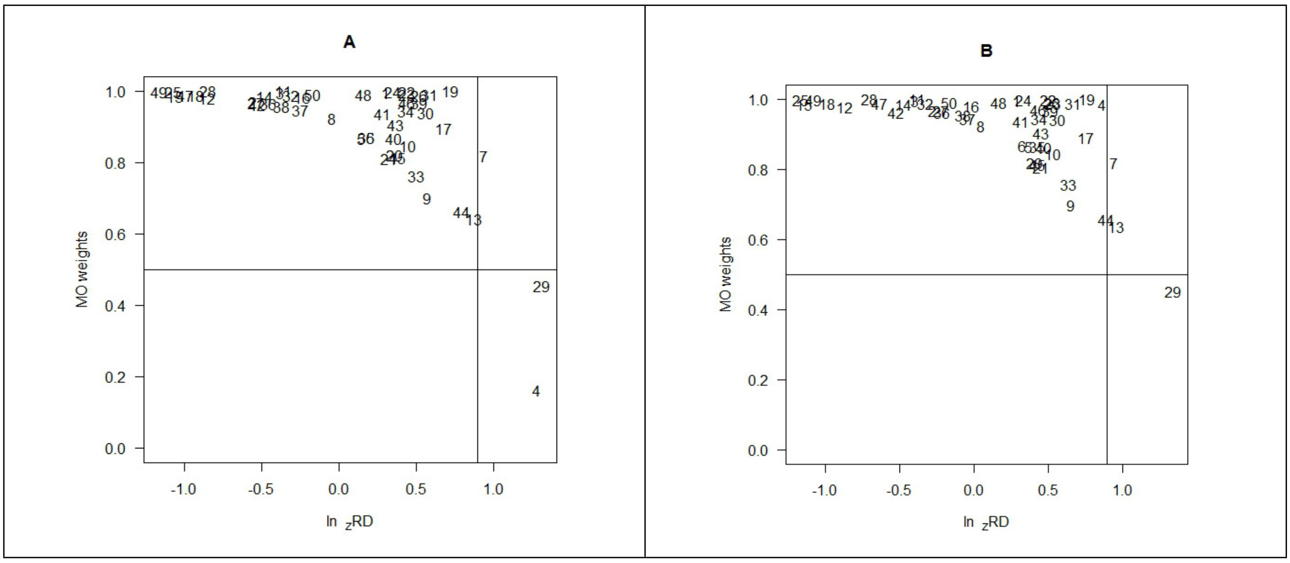


Figure 5. MO distribution of the weights given to observations in four quadrants. Data with the outlier (obs. n.4) (Panel A). Data without the outlier (Panel B)

CONCLUSIVE REMARKS

We have shown how the heteroskedastic pattern of the residuals leads to solutions that cannot be naively granted. Indeed, the Weighted Regression suggested as a not particularly sophisticated statistical method and as a standard solution to handle heteroskedasticity can be quite unsatisfactory with estimates very far from the expected ones, given the almost sure presence of outliers.

Our exemplification with data simulated according to a heteroskedastic model and with only one outlier created is not particularly illustrative of the advantages of using robust regression methods to handle potential or definite outliers, considering also that the handling of these observations is a very difficult task. However, this exemplification has shown how weighed regression even if iterated to removing the heteroskedasticity pattern can show an unsatisfactory behaviour compared to robust regression procedures.

In promoting the use of these robust procedures, special attention has been paid to the MO robust regression since its unique feature is to recover observations that are considered as outliers or at least not belonging to the bulk of observations after the first stage of its iterative process. In addition, its final figure with the residual classified as outlier or not and bad or good leverage points allows the researcher to make a sensible decision whether or not to exclude some observations from the dataset to be analysed.

A note of caution must be expressed regarding the interpretation of the determination coefficient (R^2) adjusted or not since in the context of the Weighted Regression or Robust regression it does not have the usual interpretation of the variance explained by the explanatory variables (see s.m.).

Finally, methods to deal with heteroskedasticity are not limited to weighted regression or to robust regression procedures. Indeed, since WLR estimates are as consistent and unbiased as those from OLS regression as long as the mean function in the regression model is correctly specified, it is possible to focus on the variability and to adopt methods that lead to bootstrapped standard errors computed nonparametrically by resampling from observed data [40] or to the Sandwich Standard Errors [40, 7.2.7 Sandwich Standard Errors for Least-Squares Estimates paragraph]. However, since it is not possible to rely with certainty on the fulfilment of their assumptions and since more technical statistical knowledge is required, it is strongly recommended to rely on the robust regression models and in particular on robust MO regression.

No funds have to be acknowledged.
No conflicts of interest have to be declared.
No acknowledgements have to be expressed.

REFERENCES

1. Atkinson A, Riani M. Robust Diagnostic Regression Analysis Springer 2000 New York
2. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection 1987 and 2003, John Wiley & Sons Inc.
3. Maronna RA, Martin RG, Yohai VJ. Robust Statistics: Theory and Methods 2006 John Wiley & Sons, Ltd. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.
4. Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M. Eds Robust Statistics: Theory and Methods (with R) 2019 John Wiley & Sons Ltd.
5. Huber PJ. Robust Statistics 2005 John Wiley & Sons, Inc.
6. Huber PJ. Ronchetti EM. Robust Statistics. 2nd Ed. 2009 John Wiley & Sons, Inc.
7. Draper NR, Smith H. Applied Regression Analysis 3rd Ed. 1998 by John Wiley & Sons, Inc.) (Draper NR, Smith H. Applied Regression Analysis 3rd Ed. 1998 John Wiley & Sons, Inc. NY USA)
8. Kutner MH, Nachtsheim CJ., Neter J, Li W. Applied Linear Statistical Models 5th Edition-McGraw-Hill Irwin Companies, Inc., New York, NY, 2005, 1996, 1990, 1983, 1974 Pag 426-
9. Chatterjee S, Hadi AS. Regression analysis by example 2012 5th Ed. John Wiley & Hoboken, New Jersey. Chatterjee S, Price B. 1977, 2nd Ed. 1991, 3rd Ed. 1999, 4th Ed. 2006. Hadi AS, Chatterjee S. Regression Analysis by Example Using R. Sixth Ed. 2023
10. Hoaglin DC, Welsch RE. (1978). The hat matrix in regression and ANOVA. The American Statistician, 32(1), 17-22. doi: 10.2307/2683469
11. Orenti A, Marano G, Boracchi P, Marubini E. Pinpointing outliers in experimental data: the Hat matrix in Anova for fixed and mixed effects models JPH 2012, 9, 4 e 8663 – 1-13
12. Carroll RJ, Ruppert D. Transformation and Weighting in Regression Chapman and All NY 1988.
13. Rousseeuw PJ, Yohai, V. J. (1984). Robust Regression by Means of S-estimators. Robust and Nonlinear Time series, J. Franke, W. Härdle and R. D. Martin (eds.), Lectures Notes in Statistics 26, 256-272, New York: Springer.
14. Hoaglin DC, Welsch RE. (1978), The hat matrix in regression and ANOVA, Am. Stat., 32, 17-22.
15. Henderson HV, Velleman PF. (1981), Building multiple regression models interactively, Biometrics, 37, 391-411.
16. Cook RD, Weisberg S. (1982), Residuals and Influence in Regression, Chapman & Hall, London
17. Hoaglin DC, Mosteller F, Tukey JW. (1983), Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, New York.
18. Paul SR. (1983), Sequential detection of unusual points in regression, The Statistician, 32, 417-424.
19. Stevens JP. (1984), Outliers and influential data points in regression analysis, Psychol. Bull., 95, 334-344.
20. Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. (1980), John Wiley & Sons, New York.)
21. Cook RD, 1977 Detection of influential observation in linear regression, Technometrics 19, 15-18.
22. Draper NR, John JA. 1981. Influential observations and outliers in regression. Technometrics 23, 21-26
23. Atkinson AC, 1982 Regression diagnostics, transformations and constructed variables. J. R. Stat. Soc. Ser. B, 44, 1-36
24. Hocking RR. (1983) Development in linear regression methodology: 1959-1982, Technometrics 25, 219-249
25. Hocking RR, Pendleton OJ. (1983) The regression dilemma. Commun Stat (theory and Methods) 12, 497-527
26. SAS Institute Inc. 2016. SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc
27. Atkinson AC. (1983), Diagnostic regression for shifted power transformations, Technometrics, 25, 23-33
28. Velleman PF, Welsch RE. (1981), Efficient computing of regression diagnostics, Am. Stat., 35, 234-242.
29. Montgomery D C, Peck AE. Introduction to Linear Regression Analysis (1982), John Wiley & Sons, New York.
30. Nagelkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. Biometrika 78, 691–692.
31. Singer JD, Willett JB. (2003) Applied Longitudinal Data Analysis - Modeling Change and Event Occurrence. University Press Scholarship Online Oxford Scholarship Online.
32. Cohen J, Cohen P, West SG, Aiken LS. (2002) Applied multiple correlation/regression analysis for the social sciences third Ed. Lawrence Erlbaum Associates, Publishers 2003 Mahwah, New Jersey London)
33. Mosteller F, Tukey JW. (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley Publishing Company Reading, MA USA.
34. https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_nlin_examples02.htm.
35. Holland PW, Welsch RE. (1977) Robust regression using iteratively reweighted least-squares Communications in Statistics - Theory and Methods Communications in Statistics - Theory and Methods 6 (9), 813-827.
36. Street JO, Carroll RJ, Ruppert D. (1988). A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares. The American Statistician, 42(2), 152–154.
37. Heiberger RM, Becker RA. (1992). Design of a Function for Robust Regression Using Iteratively Reweighted Least Squares. Journal of Computational and Graphical Statistics, 1(3), 181–196.
38. Green PJ. (1984) Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives Journal of the Royal Statistical Society. Series B (Methodological) 46 (2), 149-192.
39. Orenti A, Marubini E. Robust regression analysis: a useful two stage procedure, Communications in Statistics - Simulation and Computation, 2021;50:16-37, doi: 10.1080/03610918.2018.1547400
40. Fox J, Weisberg S. An R Companion to Applied Regression 3rd Ed. 2019 Sage Publications, Inc 2455 Teller Road Thousand Oaks, California.

DATA SET: 50 observations with one created outlier (type: outlier = 0 / bulk = 1 observation n.4). The data have been simulated, according to a heteroscedastic pattern, starting from the values (means and standard deviations rounded) of the Troponin (X, mg/L) and the weight (Y, g) of the infarcted myocardial area of beagle dogs obtained by heart dissection after being sacrificed shown in the book from: Cesana Bruno Mario, Antonelli Paolo and Pea Giuseppe. La Statistica per le Scienze Biomediche 2012 Libreria Universitaria.
Particularly, the 50 data (rounded to the second decimal figure) have been simulated according to a Gaussian distribution with mean equal to 14 (μ_X) and standard deviation equal ($\sigma_X = 6$).
In addition, the regression parameters are: slope (β_1) = 0.9 and the intercept (β_0) = 0.9.
Furthermore, a heteroskedastic error has been added by increasing the measurement error of the Y variable (created as the dependent variable of a straight linear regression) by an increasing quantity depending on the increase of the X variable.

R code for the simulation process of X and Y variables with heteroskedastic errors

```
# simulation of the X values;
set.seed(seed = 135791)
X_Trop=round(rnorm(n,muX,sigmaX),2)

# fixing the standard error of the heteroskedastic residuals;
sigma_err = 6.84

# simulation of errors according to a Gaussian distribution;
set.seed(seed = 24682)
err_sim=rnorm(n,mean = 0,sd =  sigma_err)

# calculation of heteroskedastic errors proportional to X variable;
hmin=1; hmax=4
# hmax and hmin are multiplicative factors of the error_Xmax and of the error_Xmin
# multiplicative (linear) factor equal to 1 for Xmin and equal to 4 for Xmax
# the following equation is the equation of a straight line passing through 1 and 4;
fm=sqrt(((hmax-hmin)*(X_Trop-min(X_Trop))/(max(X_Trop)-min(X_Trop)))+hmin)

# calculation of the heteroskedastic errors;
err_etsc= err_sim*fm

# calculation of the values of Y_Weight as the dependent variable of a linear regression
Y_Weight = round(beta0 + beta1*X_Trop + err_etsc,2)
dt_TropWeight = as.data.frame(cbind(X_Trop,Y_Weight))

# creation of only one outlier: observation n.4
out_indx= 4
outls= c( 1.59, 15.09)

# substitution of the observation n.4° in the data set dt_TropWeight with the outlier (outls);
dt_TropWeight_with_1outls= dt_TropWeigh
dt_TropWeight_with_1outls[out_indx,1:2] = outls
```

According to a SAS ® code for reading the data:

```
DATA Trop_weigth_outlier;
INPUT NUM X_Trop Y_Weight TYPE @@; *TYPE = OUTLIER = 0 / BULK = 1;
CARDS;
1   21.09  20.25    1  2   13.61  14.63    1
3   18.03  15.60    1  4    1.59  15.09    0
```

```
5    12.78  16.35    1  6    19.14  12.83    1
7    23.64  27.59    1  8    18.27  13.17    1
9    14.05   5.33    1  10   17.17  20.92    1
11   17.74  16.52    1  12   15.49  12.24    1
13   18.46  25.17    1  14   17.54  14.57    1
15   14.56  11.61    1  16    9.80  10.46    1
17   25.07  18.98    1  18   13.06  10.44    1
19    2.81   1.92    1  20   12.70   5.82    1
21   18.53  11.21    1  22    5.46   4.87    1
23    5.26   5.63    1  24   21.45  20.49    1
25   16.06  14.16    1  26    5.55   3.23    1
27   12.48  13.57    1  28   11.71  10.48    1
29   25.00  33.71    1  30   21.42  22.98    1
31    3.87   4.38    1  32   11.00   8.35    1
33   20.35  12.21    1  34   20.15  21.69    1
35   10.42  14.07    1  36   13.97  15.16    1
37   14.28  15.98    1  38   11.96  13.46    1
39    6.21   2.89    1  40   20.92  14.46    1
41   18.52  20.33    1  42   14.14  10.26    1
43   21.71  16.02    1  44   17.22  23.71    1
45   12.72   5.73    1  46    6.94   3.63    1
47   13.54  13.71    1  48    8.34   5.96    1
49   12.55  10.86    1  50   10.14   7.76    1
; RUN;
```

It has to be noted that the observation 4 (highlighted in bold) has been created as an outlier in the Y-direction by a huge increase of its ordinate (from 2.33 to 15.09) and keeping the original abscissa value from the simulation process. So, the dataset without the outlier is obtained by inserting the Y-value of 2.33 instead of 15.9 in observation n.4.

Linear Regression - Statistical Theory

The simple regression model: $Y_i = \alpha + \beta X_i + \varepsilon_i$ can be more conveniently written in matrix terms allowing for immediate extension to multiple regression. Consequently, the two parameter estimates in the case of a simple regression (a: intercept and b: regression coefficient/slope) are incorporated into a vector **b** (2 rows and 1 column) and, accordingly, they will be defined as b_0 and b_1, respectively, or as b_0 to b_k parameters estimates in the case of a multiple linear regression with k independent variables. So, in a sample of size n, the model pertinent to the i-th observation (case) for a linear regression and also for a general linear model is:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \text{ with } i = 1, ..., n \qquad (1)$$

where: y_i is the random dependent variable (response); **x'**$_i$ is the i-th row of the matrix **X** of size n x (k+1). In observational studies aimed to evaluate the role of the independent variables in explaining the response we can think of $(\mathbf{x}_i', y_i)$ as a point in a (k+1) dimensional space. On the contrary, in experimental studies **X** is a non-random matrix defined by the structure of the experimental design and, consequently it is predetermined by the experimenter [10-11]. In both settings, the rank of the matrix **X** is the number of its independent column equal to k+1 with the first column consisting of all 1 for obtaining the intercept of the regression analysis or the grand mean of the experimental design. Of course, in the WLS regression with observation weights other than 1, the first column will consist of the actual values of the weights.

β is the (k+1) x 1 row vector of the parameters to be estimated with b_0 and b_1 usually used to indicate the intercept and the slope of the regression model, respectively; ε_i is a random error assumed to be identically, independently normally distributed (i.i.d.) with mean vector 0 and constant variance σ_ε^2;

obviously, the vector **ε** will be multivariate normally distributed with mean vector 0 and a diagonal variance-covariance matrix nxn $\mathbf{I}\,\sigma_{\varepsilon}^{2}$.
Furthermore, the residual (ri) is defined as the difference between the observed value of the dependent variable (yi) and the calculated/fitted y value defined $\hat{y}_i$ (read as y-hat); in particular: $r_i = y_i - \hat{y}_i$. So ri corresponds to the estimate of εi when the equation is written with the estimates b0 and b1 instead of their parameters. In fact, $Y_i = \alpha + \beta X_i + \varepsilon_i \rightarrow \varepsilon_i = Y_i - \alpha + \beta X_i$ which corresponds in the sample as $r_i = Y_i - (b_0 + b_1 X_i)$ and $\hat{Y}_i = b_0 + b_1 X_i$. Of course, a corresponds to b0 and b to b1.

Furthermore, it should be noted that the variance of the error term εi, defined $\sigma_{\varepsilon_i}^{2}$, is equal to: $\sigma_{\varepsilon_i}^{2} = E\left\{\varepsilon_i^{2}\right\} - \left(E\left\{\varepsilon_i\right\}\right)^{2}$. Since the expected value of εi (E{(εi}) is equal to 0, according to the assumptions of the regression model, the expected value of $\sigma_{\varepsilon_i}^{2}$ $(E\left\{\varepsilon_i^{2}\right\})$ is right $\sigma_{\varepsilon_i}^{2}$ with the conclusion that the squared residual (r_i^{2}) is an estimator of the error variance and its absolute value (|ri|) is an estimator of its standard deviation obtained by the squared root of the variance. Of course, in the case of homoskedasticity, since the residual variances are all equal, $\sigma_{\varepsilon_i}^{2}$ can be replaced by σ^{2} as the parameter of the error variance estimated by: $s^{2} = \frac{1}{n-p}\sum_{i=1}^{n} r_i^{2}$.
Thus, it is very easy to conclude that the residuals (squared or absolute value) can be used to assess the relationship between the error variance (standard deviation) function with the pertinent independent or dependent or fitted variables in order to assess the presence of a heteroskedasticity pattern and ultimately model it.
For this purpose, it is also necessary to emphasize that in the presence of potential outliers in the dataset, regressing the standard deviation function should be preferred since it is less affected by the presence of outliers than regressing the squared residual function regression which is affected by the higher squared residuals. This very important point should be kept in mind in order to understand the role of the residuals in creating the weights for the WLS-R.
For easy reading, we report the formulas of the sample slope (b) and intercept (a) estimates from the ordinary least squares (OLS):

$$b = \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} X_i^{2} - n\bar{X}^{2}} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^{2}};\ a = \bar{Y} - b\bar{X}$$

Regression analysis with the outlier

Table S1 shows the ANOVA table of the OLS regression analysis. The formulas are shown along with the actual values obtained from the OLS regression with the dataset shown above.

Table S1: ANOVA table of the simple linear OLS regression.

Source of variability	Degrees of freedom	SS	MS	Statistics F P	E(MS)
Regression Due to b1|b0	1	$\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2$ 1363.18469	$b_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2$ 1363.18469	69.63 <0.0001	$\beta_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2+\sigma^2$
Residual (error variance) MSE	n – 2 50-2=48	$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$ 939.89224	$s^2=\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{(n-2)}$ 19.58109		σ^2
Total (corrected)	n – 1 50-1=49	$\sum_{i=1}^{n}y_i^2-n\bar{y}^2$ 2303.07693	-		-
Correction term Due to b0 (a – intercept))	1	$n\bar{y}^2$ 8989.79587	$n\bar{y}^2$ 8989.79587		$n\left(\sigma^2/n+\left(\alpha+\beta\bar{X}\right)^2\right)=$ $\sigma^2+n\left(\alpha+\beta\bar{X}\right)^2$
Total	N 50	$\sum_{i=1}^{n}y_i^2$ 11292.87280			

In addition, the estimated values of the intercept and slope along with their standard errors, t statistics, and significance level are reported in the Table S1.1.

Table S1.1

Estimates	Value	Standard error	Statistics t	p-value
Intercept (b0)	0.42970	1.67672	0.26	0.7988
Slope (b1)	0.90325	0.10825	8.34	<0.0001

Furthermore, the Square Root of the MSE (Mean Square Error) is equal to 4.42505 from √19.58109. Finally, the $R^2 = 0.5919$ and the adjusted R^2 (Adj R^2) = 0.5834;

Outlier diagnostics: theory

Of course, outliers can be in the Y-direction or in the X-direction; the latter are usually called “leverage points” to be further defined as “good” or “bad leverage points”. Good leverage points are observations that are far from the bulk of the observations (observations close to the regression line determined by most of the data), but with no or almost irrelevant influence on the estimates. Otherwise, “bad leverage points” are far from the regression line determined by most of the data and can even have a dramatic impact on the estimates by pulling the regression line towards themselves. Often these “bad leverage points” are also outliers.

First, we need to consider the diagnostics based on the residuals obtained from OLS, but it should be emphasized that the OLS estimator has a very low performance (robustness) since only one observation (outlier) is sufficient to obtain parameter estimates that are much biased compared to those obtained only on the “bulk” of the observations. Furthermore, since the “breakdown point” in the one-dimensional estimation of location defined as the “smallest fraction of contamination (outlier observations) that can cause the estimator to take values arbitrarily far from the values estimated without any contamination”, the OLS estimator has a breakdown point equal to 1/n, which tends to zero when the sample size n is getting large [13]. So, it is possible to say that the estimator "breaks down" leading to the "breakdown point" expression.

First of all, it has to be remembered the "leverage h_{ii}" as the value on the diagonal of the so-called "hat matrix" **H** is defined. The name is due to the fact that this matrix transforms the observed vector **y** into its OLS estimates and then it is like putting a "hat" on the **y** vector: $\hat{\mathbf{y}} = \mathbf{Hy}$ and the "hat" is the symbol for "estimate". The hat matrix **H** is idempotent (**HH** = **H**) and symmetric (**H'=H**) and is given by: $\mathbf{H} = \mathbf{X(X'X)^{-1}X'}$.

Most authors such as Hoaglin and Welsch [14], Henderson and Velleman [15], Cook and Weisberg [16], Hocking et al. [17], Paul [18], Stevens [19], and Belsley et al. [20] determine potentially influential points by looking at the "h", and paying particular attention to points for which $h_{ii},>2p/n$ even if some people recommend a more conservative cut-off value of 3p/n, being "n" the number of observations used to fit the model and "p" the number of the parameters.

However, the hat matrix **H** completely neglects outliers in the Y-direction since **H** is based only on the X variables. Furthermore, the h_{ii} diagnostics are completely vulnerable to the "masking effect" consisting in the fact that one outlier masks another outlier. Masking occurs when a cluster of outlying observations skews the mean and the covariance estimates toward it, and the resulting distance of the outlying point from the mean is small; thus, real outliers in the X-direction can be masked by the effect of "good leverage points" that attract the regression line.

Another measure of leverage is the Mahalanobis distance or better Mahalanobis distance squared (MD_i^2) that should point to observations for which the explanatory part is far from the bulk of the data and which has a one-to-one relationship with the diagonal elements of the **H-**hat matrix. The MD_i^2 values can be compared with the 95% quantiles of the χ^2 distribution with k+1 degrees of freedom.

Moreover, other types of residual diagnostic can be computer, namely the "standardized residuals" (r_i/s, where s^2 is an unbiased estimator of s^2 when the measurement errors are independent and normally distributed with zero mean and standard deviation s), the "studentized residuals" or the "internally studentized residual" [$t_i = r_i/(s\sqrt{}(1-h_{ii})$] recommended by Hoaglin and Welsch [14], Cook and Weisberg [16], Paul [18], Stevens [19] Cook [21], Draper and John [22], Atkinson [23], Hocking [24], and Hocking and Pendleton [25].

It has to be stressed that it is possible to find a confusing denomination in the literature since the "studentized residuals" are sometimes called "standardized residuals". Finally, the term "studentized residual" is mostly applied to the studentized residuals obtained by: [$t_{(i)} = r_i/(s_{(i)}\sqrt{}(1-h_{ii})$] where $s_{(i)}$ is the estimate of σ from the regression carried out without the i-th case that are also called "*studentized deleted residuals*" or "*externally studentized residuals*" or also, according to Rousseeuw and Leroy (1983) [2] "*jakknifed residuals*" from the jackknife estimator technique of a parameter in which one systematically excludes one observation at a time from a data set, calculating the parameter estimate on the remaining observations, and then aggregating these calculated estimates. SAS® [26] calls these residuals "*RSTUDENT*" according to Belsey et al. [20] and reports that "The "RSTUDENT" residual differs slightly from "internally studentized residual" (called "STUDENT") since the error variance is estimated from $s_{(i)}^2$ without the i-th observation, not from s^2(calculated on all the observations)". Atkinson [27] referred to t(i) as *"cross-validatory residual*". Finally, Cook and Weisberg [16], and Velleman and Welsch [28] call t_i, an *"internally studentized residual"* and $t_{(i)}$ an *"externally studentized residual"* definitions that besides "jakknifed residuals" are, in our opinion, the more shareable and used in statistical jargon.

Observations with "RSTUDENT" larger than 2 in absolute value may require some attention and observations with an "internally studentized residual" greater than 3 (in absolute value) are generally considered outliers.

Assuming that the residuals are normally distributed and considering that the "studentized residuals" are practically equivalent to the "standardized residuals" when the sample size is more than 30, it is possible to say that they follow a standardized normal distribution with mean zero and variance equal to 1. Then, the probability of having a residual outside ±2 (or 2 in absolute value) is about 0.05, a value that can be considered too high and, consequently, it has to conservatively increased to ±2.5 for decreasing the probability to about 0.01 or even to ±3 for having a probability value of about 0.0026. In fact, it has to be noted that there is a conservative approach in declaring an observation as an outlier to be perhaps deleted from the original dataset.

Taking into account that the influence of the i-th observation can be asserted by considering the results of the regression carried out both with and without that observation, some so-called "single-case diagnostics" have been proposed. Particularly, the Cook's squared distance [21] measures the change in the regression coefficients that would occur if a case was omitted.

Cook and Weisberg [16] and Montgomery and Peck [29] suggested that a value around 1.0 deserve attention since it is generally considered large.

A rule of thumb is that any observation with a Cook's distance greater than 4/n (where *n* is the number of the observations) is considered highly influential on the regression estimates and, consequently, should be considered as a potential outlier capable of biasing them.

Furthermore, Belsey et al. [20] proposed the DFFITS(i) statistics (very similar to Cook's distance) as a measure of influence on the prediction with a potential alarming threshold over 2√(p/n) and the DFBETASj(i) statistics, that are a scaled measure of the change in each parameter estimate (the j-th regression coefficient) calculated by deleting the i-th observation from the dataset with a cut-off value of 2/√n (better than just 2 without considering the sample size).

Another statistic to be considered is the "COVRATIO" that measures the change in the determinant of the covariance matrix estimated by deleting the i-th observation. Belsey et al. [20] suggest that an absolute value of (COVRATIO - 1) ≥ 3p/n deserves to be investigated as a potential outlier. Actually, the COVRATIO is the ratio between the determinant of the variance covariance matrix without the i-th observation and the determinant of the variance covariance matrix with the i-th observation included.

However, according to Rousseeuw and Leroy [2] as well, all these diagnostics have an interpretation no longer reliable when the data contain more than one outlier, and are susceptible to the masking effect with the unfortunate conclusion that they often fail to identify outliers.

Furthermore, we must remember the extension of most single-case diagnostics to multiple-case diagnostics with the Cook distance generalized by Cook and Weiberg [16] being the most relevant.

Finally, the "Resistant Diagnostic" has been proposed by Rousseeuw and Leroy [2, pages. 238-245) to which the interested readers are referred.

Outlier diagnostics: *data with the outlier*

The OLS residual statistics allow us to consider as potential outliers seven observations.

Particularly the residual statistics of the observation n.4 passed seven thresholds (leverage, externally studentized residuals, internally studentized residuals, Cook's distance, DFFITS, DBETAS intercept, and COVRATIO) but the residual statistics of the observation n. 29 also passed six thresholds (leverage, externally studentized residuals, Cook's distance, DFFITS, DBETAS slope, and COVRATIO). Then the residual statistics of the observations n. 19 and 31 passed two thresholds (leverage and COVRATIO); finally, the residual statistics of the observations n. 7, n. 17, and n. 23 passed only one threshold: DBETAS slope, leverage and COVRATIO, respectively.

Thus, it is almost obvious that the observation n. 4, modified to be an outlier in the Y-distance, has OLS residual statistics greater than too many thresholds causing it to be confirmed as a "true outlier".

Figure S1.1 and S1.2 show some plots of the OLS regression obtained from SAS® Proc REG with the keyword "influence".

On the left (Figure S1.1), there is the distribution of the studentized residuals (internally studentized residuals) with two vertical lines drawn at ±3 as the thresholds for considering the corresponding observation as an outlier and the Cook's distance with its threshold (see comments of the Figure 5) for considering the corresponding observation as a high leverage point. On the right, there is the diagram plot of the Y and X variables together with the regression line and the pointwise 95%CI of the fitted values (the two concave and convex lines delimit the black space near the line; as expected the point with the Y and X means as coordinates has the narrowest interval). In addition, there are the

95% prediction limits (95% CI of a generic Y value at the corresponding abscissa) shown as dotted lines that are apparently parallel owing to a minimum expected concave-convex pattern with the narrowest interval at the point with the means of Y and X as coordinates. It should be emphasized that the confidence probability of 95% has to be referred to each of these 95%CIs. To refer the confidence probability of 95% to all the intervals, the simultaneous 95%CIs must be calculated.

Figure S1.1

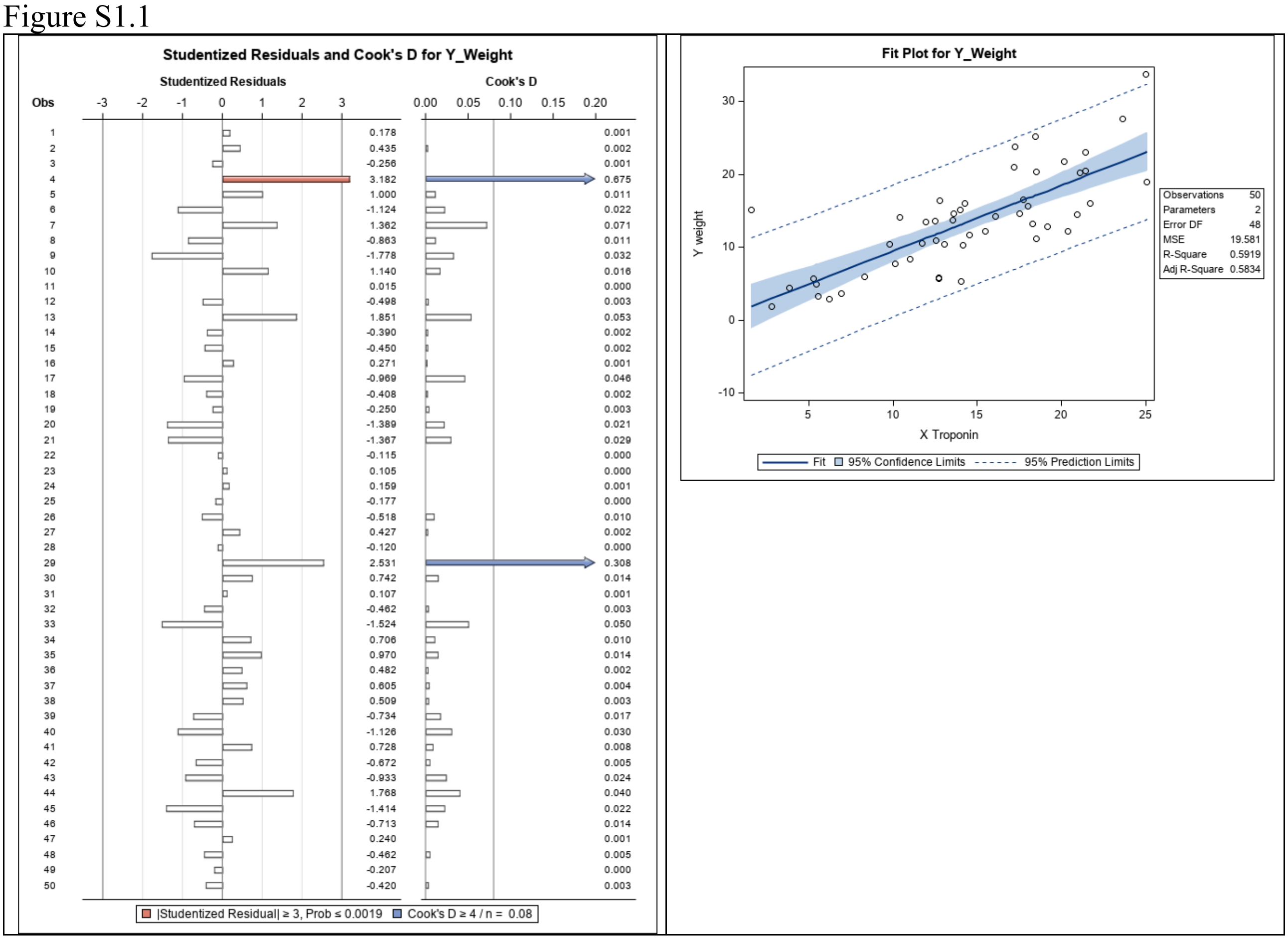
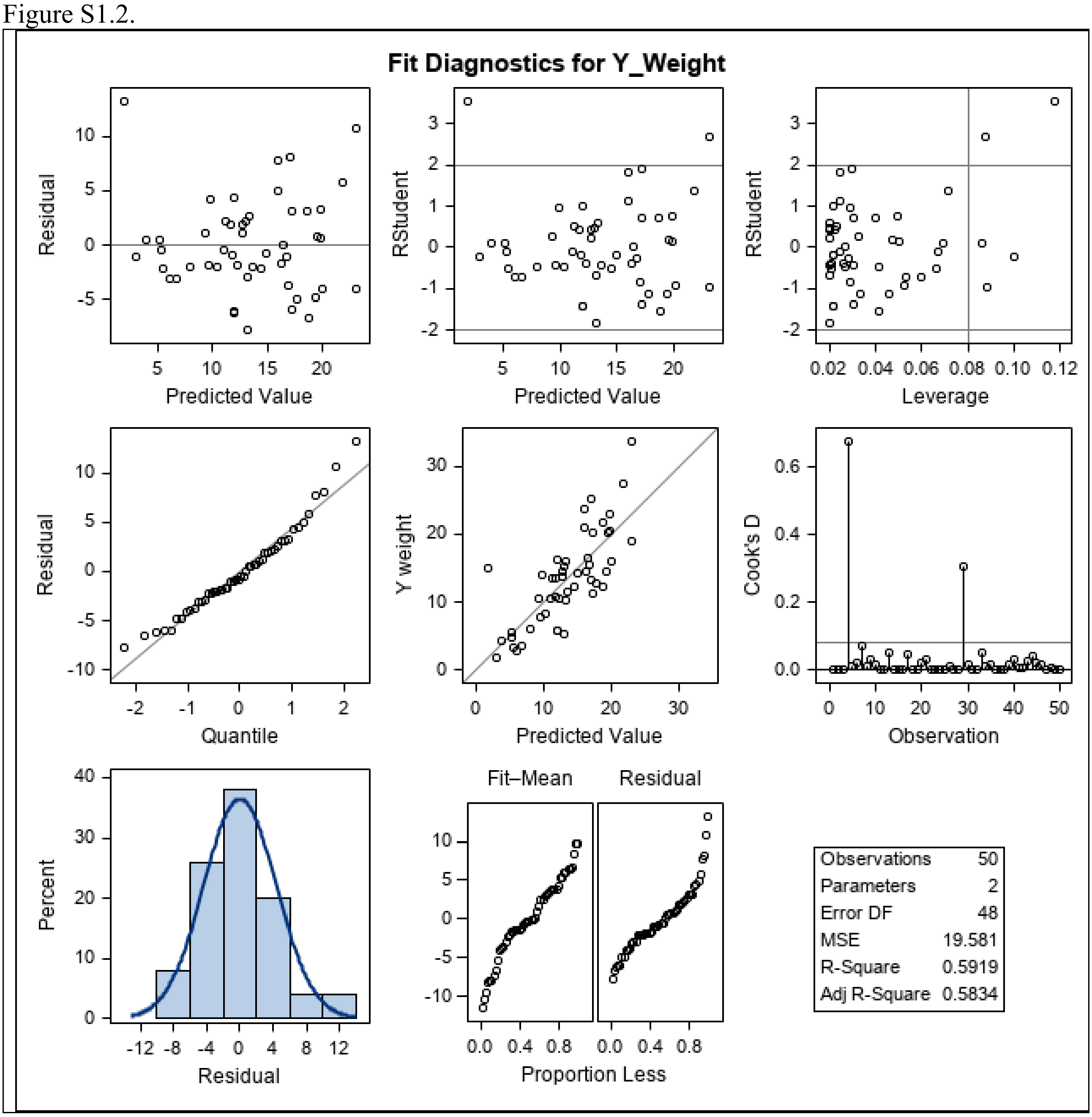


Figure S1.2.



Fit Diagnostics for Y_Weight

Observations	50
Parameters	2
Error DF	48
MSE	19.581
R-Square	0.5919
Adj R-Square	0.5834

Figure S1.2 shows:

A.1)-Top row:

A.1.1)-left: plot of the "residuals" vs. the "predicted values"; the heteroskedasticity pattern is evident but it is better to consider the following plot.

A.1.2)-centre: plot of the "externally studentized residuals" vs. the "predicted values"; this plot should be considered primarily to judge the presence of a heteroskedasticity pattern. This plot corresponds to Figure 1-Panel A of the paper. The two horizontal lines at ±2 refer to the accepted thresholds to consider the "externally studentized residuals" that are inside as non-outliers;

A.1.3)-right: plot of the "externally studentized residuals" vs. the "leverage values". The two horizontal lines at ±2 refer to accepted thresholds for "externally studentized residuals"; the vertical line is drawn at the threshold of the leverage given by 2p/n equal to 4/50= 0.08.

In fact, there are two observations considered "outliers" with an externally studentized residual greater than 2 and leverage more than 0.8 (observations n.4 and n.29); then, there are three observations with leverage value greater than 0.8 (observations n.19, n.23 and n.31) and within the interval ±2.

A.2)-Centre row:

A.2.1)-left: QQ plot of the “residuals”. It seems that the residuals are Gaussian distributed since they are well superimposed on the straight line obtained according to the Gaussian distribution. Indeed, the Shapiro-Wilk test does not reject the null hypothesis of a sample randomly drawn from a Gaussian distribution (P = 0.0976; Kolmogorov-Smirnov P >0.1500; Cramer-von Mises: P >0.2500; Anderson-Darling: P = 0.2217).
A.2.2)-centre: plot of the observed Y (Y_Weight) vs. the “predicted values” with the bisector equality line of the first Cartesian quadrant that is the place where the ordinates and the abscissas are equal and of the complete agreement between two measurement methods. It is obvious that in case of a perfect regression the observed and the fitted values are equal and lie on the equality line. The poorer the relationship, the further the points will move away from the bisector line.
A.2.3)-right: plot of Cook’s distance D: it corresponds to the previous plot seen in Figure S1.1 rotated counterclockwise with a horizontal line drawn at 0.08 (threshold for this statistic to mark observations suspected to be outliers).
A.3)-Bottom row:
A.3.1)-left: histogram of the residuals overlaid with the Gaussian curve with mean and standard deviation of the sample “residuals”.
A.3.2)-centre: plot of the Fit-Mean (fitted Y predicted values) on the left and the residuals (on the right) vs. their cumulative proportion. A straight-line pattern would suggest a uniform distribution. Since a Gaussian distribution is expected, we should see an italic “S” shaped line.
A.3.3)-right: some information about the regression analysis such as the number of the observations (n), the number of the parameters (2 in the case of the simple linear regression) the degrees of freedom (n – 2 = 48, in the case of the simple linear regression) of the error variance, Mean Square Error (MSE) equal to 19.581 for the OLS regression and the values of R^2 and adjusted R^2 (0.5919 and 0.5834, respectively).

The following Figure S3 shows the plot of the Rstudent (externally studentized) residuals vs. the fitted values. In the panel A there are the Rstudent residuals after the OLS regression corresponding to the iteration 0 (“Iter 0” on the top) of the iterative process given by the IRWLS estimator.
Panel B, Panel C, and Panel D show the Rstudent residuals obtained after the first, the second and the third iteration of the weighted regression performed with the IRWLS estimator.

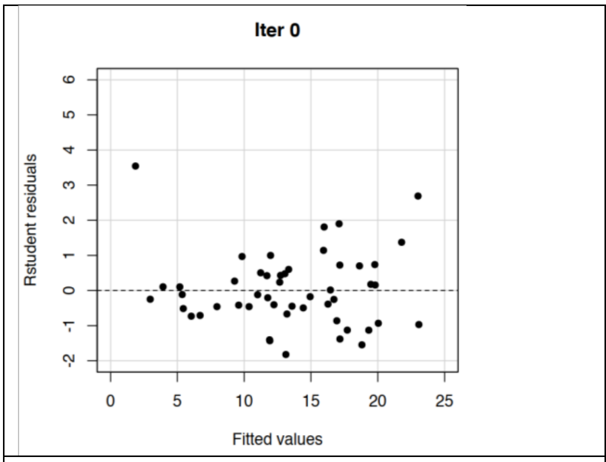
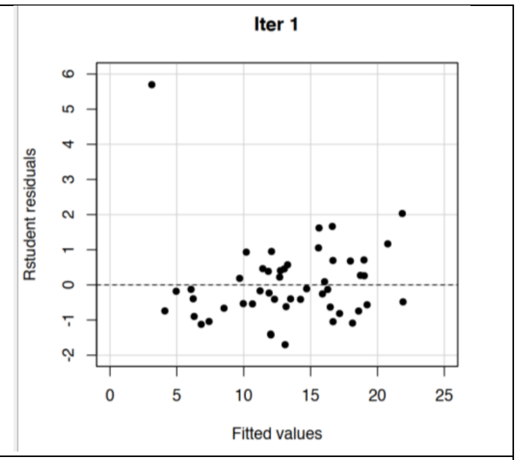
	

Figure S2-Panel A: residuals from OLS regression	Figure S2-Panel B: residuals from the first weighted regression

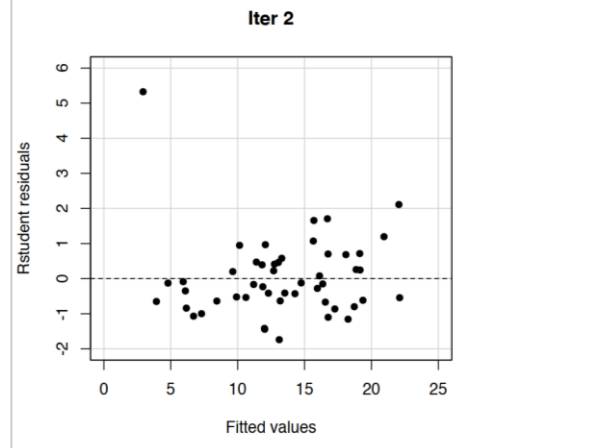


Figure S2-Panel C: residuals after the second weighted regression

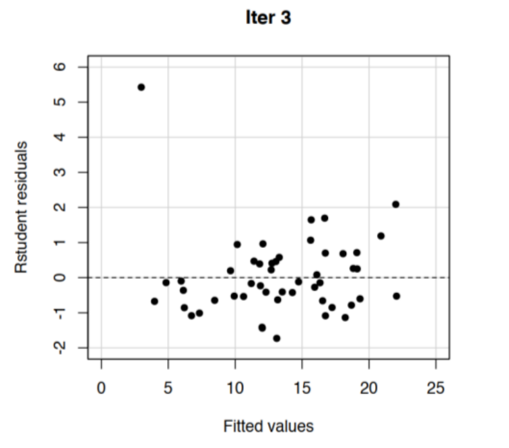


Figure S2-Panel D: residuals after the third weighted regression

It can be seen that the heteroskedastic pattern of the residuals, well evident in Figure 1-Panel A (after OLS regression), is somewhat reduced in Figure S2-Panel B with some points pushed towards the X axis. Hence, the heteroskedastic pattern practically does not change in the first and second iteration (second WS -Figure S2-Panel C- and third WS – Figure S2 – Panel D, respectively). Nevertheless, the parameter estimates of the successive iterations are somewhat different as shown in Table S2, as expected from the theoretical point of view. It should be noted that the first step shown in Table S2 corresponds to the OLS linear regression with weights equal to 1.

Table S2 – Dataset with the outlier.

Estimates	First OLS-R	Second (1WR)	Third (2WR)	Fourth (3WR)	Fifth (4WR)	Sixth (5WR)	Seventh (6WR)	Eight (7WR)
Intercept (b_0) (s.e.) Statistics t P	0.4297 1.6767 0.256 0.799	1.8534 1.4397 1.287 0.204	1.6301 1.4641 1.113 0.273	1.6912 1.4571 1.161 0.252	1.6741 1.4890 1.147 0.257	1.6789 1.4585 1.151 0.255	1.6776 1.4586 1.150 0.256	1.6779 1.4586 1.150 0.256
Slope (b_1) (s.e.) Statistics t P	0.9032 0.1083 8.344 <0.0001	0.8002 0.1058 7.562 <0.0001	0.8169 0.1055 7.740 <0.0001	0.8123 0.1056 7.693 <0.0001	0.8136 0.1056 7.707 <0.0001	0.8132 0.1056 7.703 <0.0001	0.8133 0.1056 7.704 <0.0001	0.8133 0.1056 7.704 <0.0001
MSE	4.425	4.524	4.488	4.497	4.494	4.495	4.495	4.495
R^2	0.5919	0.5346	0.5552	0.5522	0.5530	0.5528	0.5529	0.5528
Adj-R^2	0.5834	0.5431	0.5459	0.5428	0.5437	0.5435	0.5435	0.5435

From Table S2, it is possible to see that the intercept of the first WR is much greater than that of the OLS; hence, it decreases with a tendency to stabilize around values of 1.67. Otherwise, the slope of the WRs decrease to a plateau around 0.81. The standard errors of the estimates decrease during the iterative process.

It should also be noted the practically stable behavior of the MSE until a plateau around 4.495.

Considerations on the coefficient of determination

It is recalled that the R^2 and the adjusted R^2 (R^2_{adjusted}) are obtained respectively: $R^2 = 1 - \frac{(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2}$; $R^2_{\text{adjusted}} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$ where k is the number of predictors and n is the total sample size. The adjusted R^2 should be preferred for judging the goodness of fitting since it increases only when an independent variable added to the model is statistically significant and affects the dependent variable as opposed to the R^2 that increases when an independent variable is added to the model. Obviously, the adjusted R^2 value is always less than or equal to the R^2 value.

Note that the R^2 obtained by the WLS regression with the Y and X variables multiplied by their pertinent weights (**Y*** **and X***) is:

$$\mathbf{Y}_* = \mathbf{W}^{-1/2}\mathbf{Y}, \text{ and } \mathbf{X}_* = \mathbf{W}^{-1/2}\mathbf{X};\quad \mathrm{R}^2_{\mathrm{WLS}} = 1 - \frac{\left(\mathbf{Y}_* - \mathbf{X}_*\mathbf{b}_*\right)'\left(\mathbf{Y}_* - \mathbf{X}_*\mathbf{b}_*\right)}{\mathbf{Y}_*'\mathbf{Y}_* - n\bar{Y}_*^2};$$

Where **b*** is the WLS estimate of **β**. In fact, it is informative to transform the equation to create a model that can be fitted with the OLS, even though the WLS estimates are usually calculated directly. Then, multiplying throughout the usual regression equation model by the squared root of the inverse of the weight matrix (**W**$^{-1/2}$), one can obtain the above formula of the coefficient of determination ($\mathrm{R}^2_{\mathrm{WLS}}$) in the case of weighted regression.

The following formula of the coefficient of determination is not in matrix language:

$$\mathrm{R}^2_{\mathrm{WLS}} = 1 - \frac{\sum_{i=1}^{n} w_i \left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} w_i \left[Y_i - \left(\sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i \right) \right]^2}$$

Note that the above formula is essentially equivalent to the corresponding OLS formula except that instead of a weight always equal to 1, the weights have a specific value for each observation, resulting in the WLS estimator.

The weighted least squares output of some regression software packages includes R^2, the coefficient of determination (multiple in the case of more than one dependent variable). Users of these packages should treat this statistic with caution, because R^2 (adjusted R^2) does not have in weighted regression analysis the usual interpretation as in OLS regression. Indeed, the weighed R^2 is a measure of the proportion of the variation in the weighted Y than can be accounted by the weighted X. Interested readers are referred to Nagelkerke [30].

Furthermore, we agree with Willett and Singer [31] “that it is not good to rely on any R^2 (even the pseudo R$^2_{\mathrm{wls}}$) as a sole measure of goodness of fit”.

Analysis without the outlier

The OLS regression results can be seen in Table 1 and Table 2 of the paper.

Outlier diagnostics: *data without the outlier*

One threshold was exceeded by the residual statistics of observations n. 7 (DBETAS slope), observation n. 13 and observation n. 44 (Rstudent or externally studentized residuals, for both).

Two thresholds were exceeded by the residual statistics of observations n. 4 (leverage and COVRATIO), observation n. 17 (leverage and Cook’s D), observation n. 19 (leverage and COVRATIO), and observation n. 31 (leverage and COVRATIO).

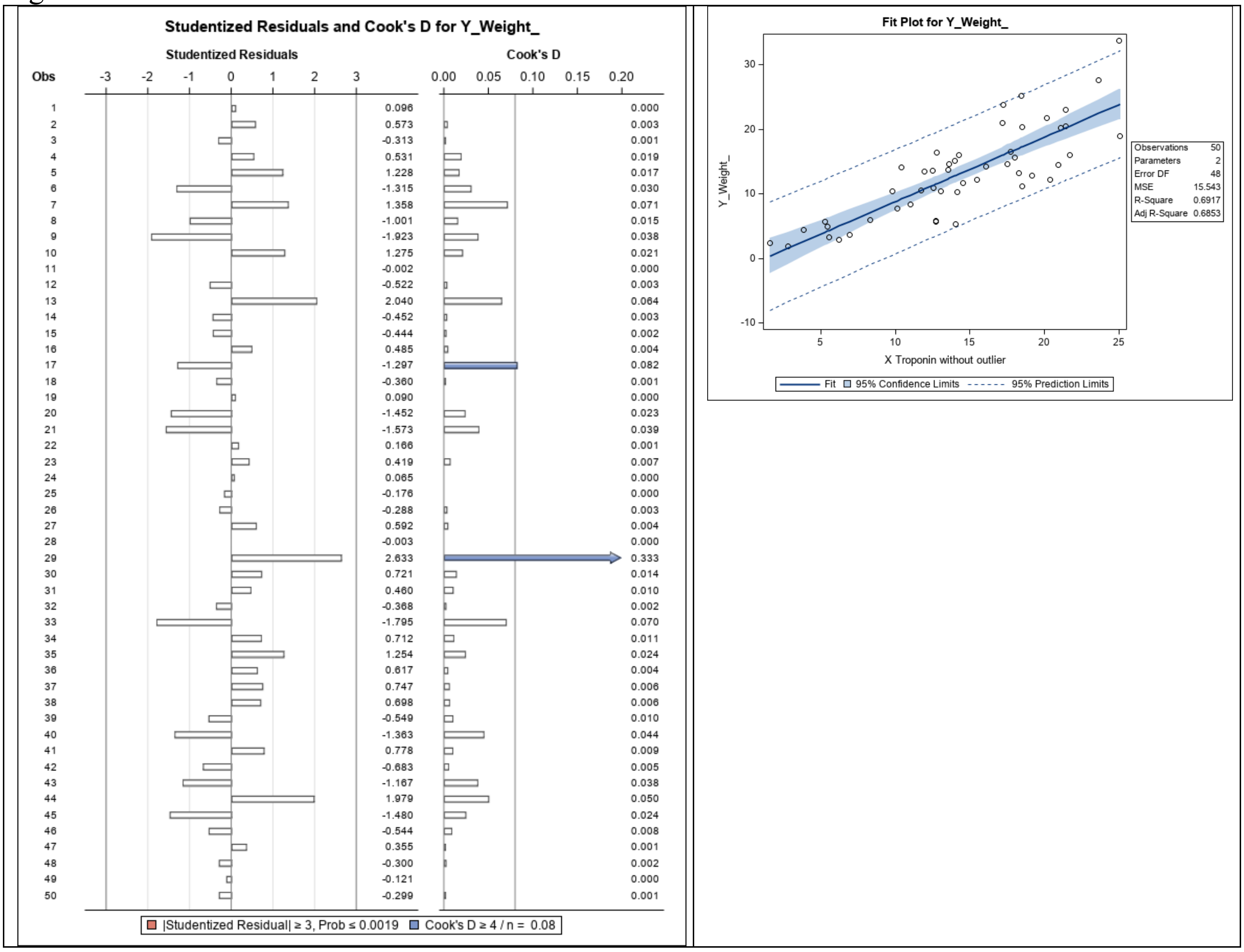
Finally, the residual statistics of observation n. 29 passed six thresholds (leverage, Rstudent or externally studentized residuals, Cook’s D, DFFITS, DBETAS slope, COVRATIO).

Regarding these data, observation n. 29 with six thresholds exceeded can be considered an outlier. It is difficult to judge the case of observations n. 4, n. 17, n. 19, and n. 31 with two thresholds exceeded.

Finally, it is possible to conclude that the observations n. 7, n. 13, and n. 44 with only one threshold exceeded can be considered as such without further action.

However, all the observations described above from the dataset with and without the created outlier, have to be checked at least for imputation errors. Finally, it is difficult and almost impossible to decide what to do with the observations with more than two exceeded thresholds since deleting these observations can be considered an arbitrary decision.

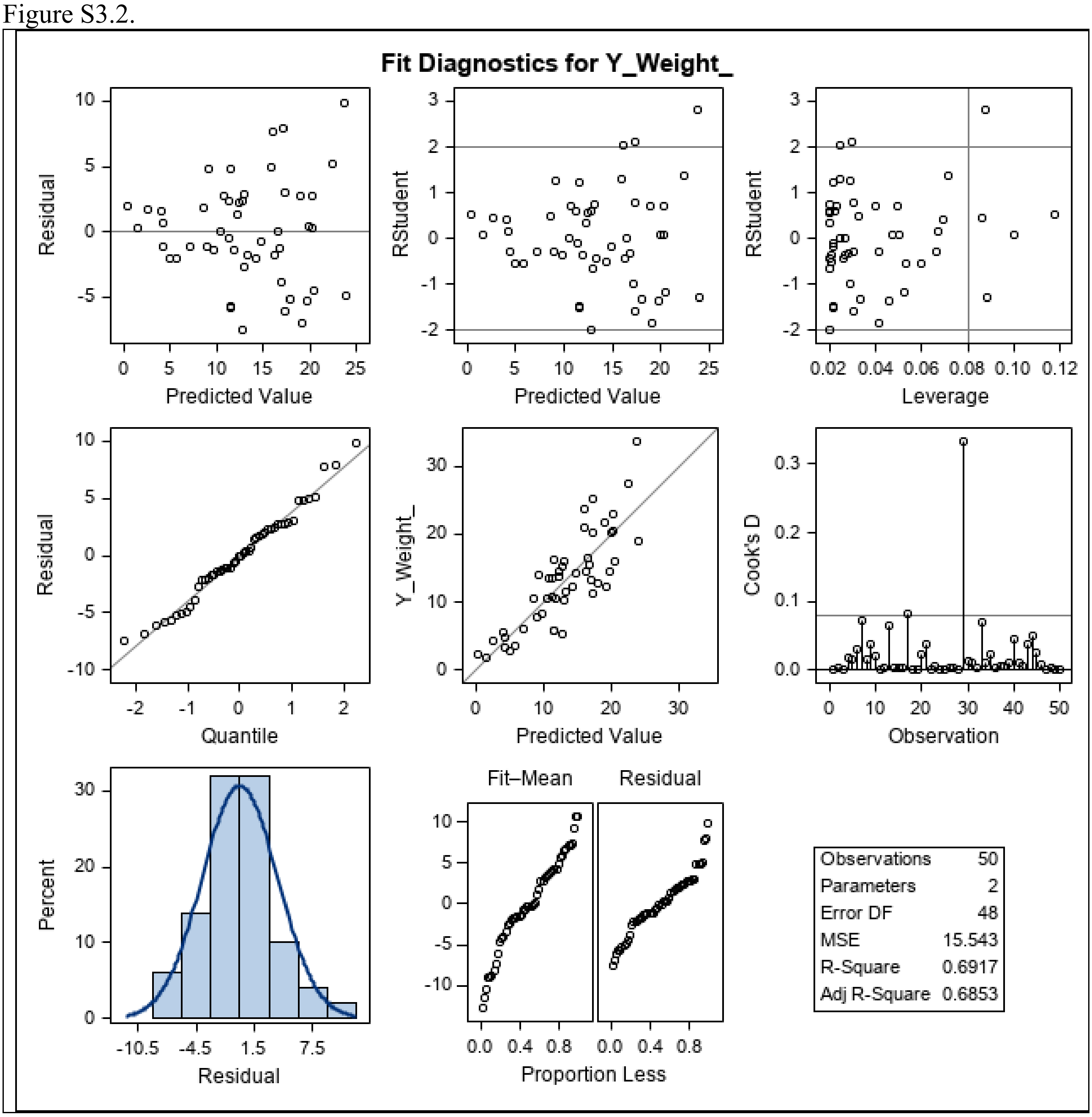
Figure S3.1



See Figure S1.1 for the comments. In this case the observation n.4 is not created as an outlier.

Figure S3.2 shows:
A.1)-Top row:
A.1.1)-left: plot of the “residuals” vs. the “predicted values”; the heteroskedasticity pattern is evident but it is better to consider the following plot.
A.1.2)-centre: plot of the “externally studentized residuals” vs. the “predicted values”; this plot is equal to the plot of Figure 1-Panel B in the paper.
A.1.3)-right: plot of the “externally studentized residuals” vs. the “leverage values”. The two horizontal lines at ±2 refer to accepted thresholds for the “externally studentized residuals”; the vertical line is drawn at the threshold of the leverage given by 2p/n equal to 4/50= 0.08.
Indeed, there is only one observation considered as “outlier” with an externally studentized residual greater than 2 and leverage more than 0.8 (observation n.29); then there are four more observations with leverage value greater than 0.8 (observations n.4, n.17, n.19, and n.31).

Figure S3.2.



Observations	50
Parameters	2
Error DF	48
MSE	15.543
R-Square	0.6917
Adj R-Square	0.6853

A.2)-Centre middle row:

A.2.1)-left: QQ plot of the “residuals”. It seems that the residuals are Gaussian distributed since they are well superimposed on the straight line obtained according to the Gaussian distribution. Indeed, the Shapiro-Wilk test does not reject the null hypothesis of a sample randomly drawn from a Gaussian distribution (P = 0.5742; Kolmogorov-Smirnov P >0.1500; Cramer-von Mises: P >0.2500; Anderson-Darling: P > 0.2500).

A.2.2)-centre: plot of the observed Y (Y_Weight) vs. the “predicted values” with the bisector equality line of the first Cartesian quadrant, i.e. the place where the ordinates and the abscissas are equal. It is obvious that in the case of a perfect regression the observed and the fitted values are equal and lie on the equality line. The poorer the relationship, the further the points will move away from the aforementioned bisector line.

A.2.3)-right: plot of Cook’s distance D: it corresponds to the previous plot in a counterclockwise direction with a horizontal line drawn at 0.08 (threshold for this statistic to mark observations suspected of being outliers).

A.3)-Bottom row:
A.3.1)-left: histogram of the residuals with the Gaussian curve with mean and standard deviation of the sample “residuals” superimposed.
A.3.2)-centre: plot of the Fit-Mean (fitted Y predicted values) on the left and the residuals (on the right) vs. their cumulative proportion “predicted values”.
A.3.3)-right: some information about the regression analysis such as the number of observations (n), the number of parameters (2 in the case of simple linear regression) the degrees of freedom of the error variance or residual variance (n – 2 = 48, in the case of simple linear regression), the Mean Square Error (MSE) equal to 15.543 for the OLS regression and the values of R^2 and of adjusted R^2 (0.6917 and 0.6853, respectively).

Figure S4 Data without the outlier. OLS “RStudent” residuals (Iter0, Panel A) and “Rstudent” residuals after the first (Panel B), second (panel C) and third (Panel D) iterated weighted regression by means of the Iterated Re-Weighted Least Squares Estimator.

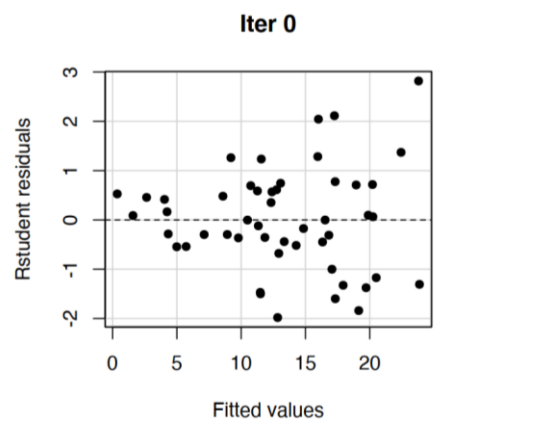
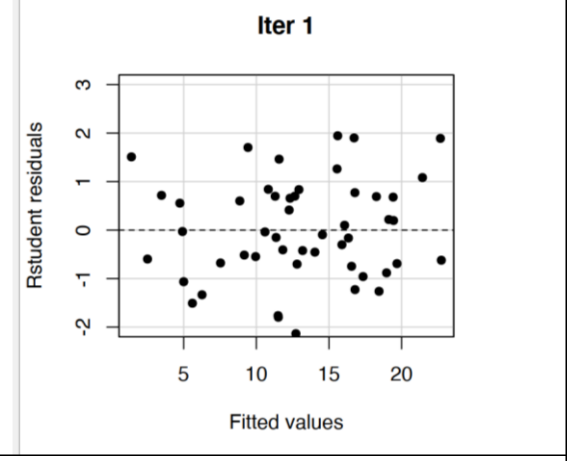
	

Figure S4.A: After OLS regression	Figure S4.B: After the first weighted regression

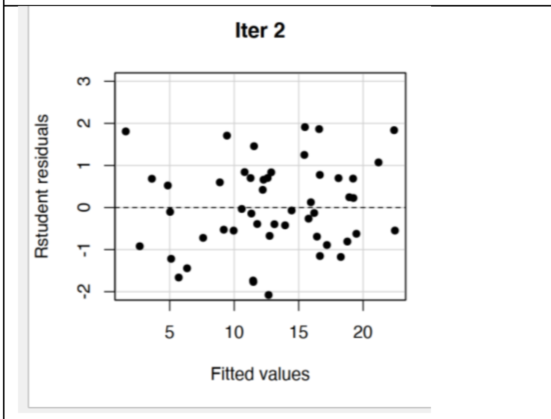
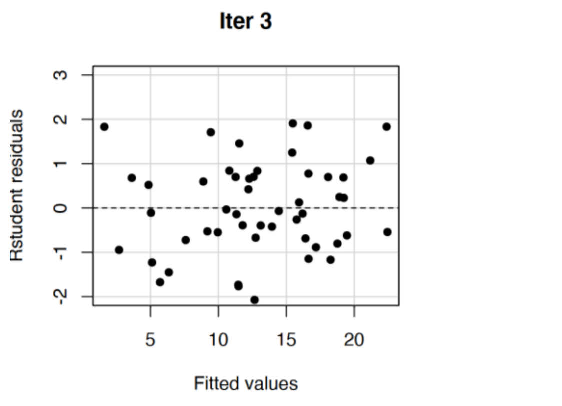

	

Figure S4.C: After the second weighted regression	Figure S4.D: After the third weighted regression

It is possible to see that the heteroskedastic pattern is practically absent since the first iteration, even if some little adjustments (the RStudent residuals tend to be more uniformly distributed) have been made at the second iteration. Finally, the “Rstudent residuals” at the third iteration are practically equal to those of the second iteration. Nevertheless, in the successive iterations the parameter estimates are somewhat different as Table S3 shows.

Table S3 – Dataset without the outlier.

Estimates	First OLS-R	Second (1WR)	Third (2WR)	Fourth (3WR)	Fifth (4WR)	Sixth (5WR)	Seventh (6WR)
Intercept (b_0) (s.e.) Statistics t P	-1.2279 1.4939 -0.8220 0.4150	-0.0423 0.3067 -0.0700 0.9450	0.1746 0.5009 0.3490 0.7290	0.1904 0.4937 0.385 0.701	0.1918 0.4931 0.389 0.699	0.1919 0.4930 0.389 0.699	0.1919 0.4930 0.389 0.699
MSE	3.942	2.596	2.309	2.287	2.285	2.285	2.285
R^2	0.6917	0.8185	0.8311	0.8319	0.832	0.832	0.832
Adj-R^2	0.6853	0.8148	0.8276	0.8284	0.8285	0.8285	0.8285

From Table S3, it can be seen that the intercept of the first WR shows a relevant increase in comparison to the OLS; then, the intercept value decreases with a tendency to stabilize around values of 0.1918. It has to recall that the fifth iteration corresponding to the fourth weighted regression. Otherwise, the slope of the WRs decreases to a plateau around 0.88 from the third iteration (second weighted regression). The standard errors of the estimates decrease during the iterative process. It should also be noted the decreasing behavior of the MSE to a plateau around 2.285. Of course, due to the absence of a relevant outlier, the iterative process is expected to converge with only a few iterations. Finally, it has to be noted that in this case the MSE decreases instead of the little increase shown for the dataset with the outlier.

Descriptive statistics of the weights given by the IRWLS, MM and LTS

Dataset without the outlier

IRWLS Method

The weights have a mean of 0.2886 (±0.06634, s.d.), median of 0.1096, first (Q_1) and third (Q_3) quartiles of 0.0651 and 0.17484, respectively, and a minimum and maximum of 0.0363 (obs. n.17) and 4.3291 (obs. n.4), respectively. No observations have a weight of 1. In addition, in this case the positive skewness of the distribution is evident since the median is much lower than the arithmetic mean. When these weights are "normalized" (divided by the sum of the weights multiplied by the number of observations) they sum equals the number of observations.

MM Method

The weights have mean of 0.9146 (±0.1159, s.d.), median of 0.9668, first (Q_1) and third (Q_3) quartiles of 0.8622 and 0.9891, respectively, a minimum and maximum of 0.4421 (obs. n.29) and 1.0000 (obs. n.28), respectively. In addition, in this case there is a distribution a little negatively skewed since the median is greater than the arithmetic mean.

LTS Method

There are 47 observations with weight equal to 1 and 3 observations equal to 0 (n.13, n.29, and n.44).

Dataset with the outlier

IRWLS Method

The weights have mean of 0.0912 (±0.03206, s.d.), median of 0.08574, first (Q_1) and third (Q_3) quartiles of 0.0688 and 0.1016, respectively, a minimum and maximum of 0.0517 (obs. n.17) and 0.1901 (obs. n.4), respectively. No observation has weight 1. In addition, in this case the distribution is almost symmetric since the median is very similar to the arithmetic mean.

MM Method

The weights have mean of 0.9069 (±0.1468, s.d.), median of 0.9675, first (Q_1) and third (Q_3) quartiles of 0.8779 and 0.9897, respectively, a minimum and maximum of 0.2000 (obs. n.4) and 0.9999 (obs.

n.28), respectively. In addition, in this case there is a distribution a little negatively skewed since the median is greater than the arithmetic mean.

LTS Method
There are 47 observations with weight equal to 1 and 3 observations equal to 0 (n.4, n.13, and n.29).

It is interesting to consider the degrees of freedom of the regression analyses with the considered 5 methods. Indeed, it has to stress that OLS, WLS with IRWLS, MO and MM have all 48 degrees of freedom for the error variance equal to n – 2 for the simple regression.
Otherwise, LTS analysis with 3 observations weighted as 0 has 45 degrees of freedom equal to 47 (the observations with weight equal to 1) minus 2.

Table S4. Basic notation for OLS and WLS method

Model	$\mathbf{y}=\mathbf{X\beta}+\ \boldsymbol{\varepsilon}$ **or** $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}^{*}\beta_j + \varepsilon_i$	
	OLS	**WLS**
Assumptions		
In observational studies only	$\mathbf{x}_i^{*} \sim \mathbf{G}_\mathbf{p}(\boldsymbol{\mu}_{\mathbf{x}^*}, \boldsymbol{\Sigma}_{\mathbf{x}^*})$	
In observational and experimental studies	$\boldsymbol{\varepsilon} \sim \mathbf{G}_\mathbf{n}\left(\mathbf{0}, \sigma^2\mathbf{I}_{(\mathbf{n})}\right) \ \rightarrow\ \mathbf{y} \sim \mathbf{G}_\mathbf{n}(\mathbf{X\beta}, \sigma^2\mathbf{I}_{(\mathbf{n})})$ $\varepsilon_i \sim G(0, \sigma^2) \quad \forall i = 1, 2, \dots, n$	$\boldsymbol{\varepsilon} \sim \mathbf{G}_\mathbf{n}\left(\mathbf{0}, {}_{\overset{\square}{w}}\sigma^2\mathbf{W}^{-1}\right) \ \rightarrow\ \mathbf{y} \sim \mathbf{G}_\mathbf{n}(\mathbf{X\beta}, {}_{\overset{\square}{w}}\sigma^2\mathbf{W}^{-1})$ $\varepsilon_i \sim G(0, \sigma_i^2) \quad \forall i = 1, 2, \dots, n$
Hat matrix	$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ $h_{ii} = x_i'(\mathbf{X}'\mathbf{X})^{-1}x_i \quad (3)$	${}_{\overset{\square}{\mathbf{w}}}\mathbf{H} = \sqrt{\mathbf{W}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\sqrt{\mathbf{W}}$ $h_{ii} = \sqrt{w_i}x_i'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}x_i\sqrt{w_i}$
Estimates		
Regression parameters	$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (4) \qquad \mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	${}_{\overset{\square}{w}}\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (4) \qquad \mathrm{Cov}\left({}_{\overset{\square}{w}}\mathbf{b}\right) = {}_{\overset{\square}{w}}\sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$
Predicted values	$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y} \qquad \mathrm{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$	${}_{\overset{\square}{w}}\hat{\mathbf{y}} = \mathbf{X}{}_{\overset{\square}{w}}\mathbf{b} = {}_{\overset{\square}{w}}\mathbf{H}\mathbf{y} \qquad \mathrm{Cov}\left({}_{\overset{\square}{w}}\hat{\mathbf{y}}\right) = {}_{\overset{\square}{w}}\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'$
Residuals	$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \left(\mathbf{I}_{(n)} - \mathbf{H}\right)\mathbf{y} \qquad \mathrm{Cov}(\mathbf{e}) = \sigma^2\left(\mathbf{I}_{(n)} - \mathbf{H}\right)$ $e_i = y_i - \hat{y}_i \qquad \mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$	${}_{\overset{\square}{w}}\mathbf{e} = \mathbf{y} - {}_{\overset{\square}{w}}\hat{\mathbf{y}} == \left(\mathbf{I}_{(n)} - {}_{\overset{\square}{w}}\mathbf{H}\right)\mathbf{y} \qquad \mathrm{Cov}({}_{\overset{\square}{w}}\mathbf{e}) = {}_{\overset{\square}{w}}\sigma^2\mathbf{W}^{-1}\left(\mathbf{I}_{(n)} - {}_{\overset{\square}{w}}\mathbf{H}\right)$ ${}_{\overset{\square}{w}}e_i = y_i - {}_{\overset{\square}{w}}\hat{y}_i \quad (5) \qquad \mathrm{Var}({}_{\overset{\square}{w}}e_i) = {}_{\overset{\square}{w}}\sigma^2 w_i^{-1}\left(1 - {}_{\overset{\square}{w}}h_{ii}\right)$ ${}_{\overset{\square}{w}}\mathbf{r} = \sqrt{\mathbf{W}}{}_{\overset{\square}{w}}\mathbf{e} \qquad \mathrm{Cov}({}_{\overset{\square}{w}}\mathbf{r}) = {}_{\overset{\square}{w}}\sigma^2\left(\mathbf{I}_{(n)} - {}_{\overset{\square}{w}}\mathbf{H}\right)$ ${}_{\overset{\square}{w}}r_i = \sqrt{w_i}{}_{\overset{\square}{w}}e_i \qquad \mathrm{Var}({}_{\overset{\square}{w}}r_i) = {}_{\overset{\square}{w}}\sigma^2\left(1 - {}_{\overset{\square}{w}}h_{ii}\right)$
Residual sum of squares	$\mathrm{RSS} = \mathbf{e}'\mathbf{e}$	$\mathrm{WRSS} = {}_{\overset{\square}{w}}\mathbf{r}'{}_{\overset{\square}{w}}\mathbf{r} = {}_{\overset{\square}{w}}\mathbf{e}'\mathbf{W}{}_{\overset{\square}{w}}\mathbf{e}$
Mean Square	$\hat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n - (p + 1)}$	${}_{\overset{\square}{w}}\hat{\sigma}^2 = \dfrac{\mathrm{WRSS}}{n - p} \quad (7)$
Diagnostics		
Scaled residual	$\dfrac{e_i}{\hat{\sigma}}$	$\dfrac{{}_{\overset{\square}{w}}e_i}{{}_{\overset{\square}{w}}\hat{\sigma}} \qquad \dfrac{{}_{\overset{\square}{w}}r_i}{{}_{\overset{\square}{w}}\hat{\sigma}}$
Standardized residual	$\dfrac{e_i}{\sqrt{(1 - h_{ii})}}$	$\dfrac{{}_{\overset{\square}{w}}e_i}{\sqrt{w_i^{-1}\left(1 - {}_{\overset{\square}{w}}h_{ii}\right)}} = \dfrac{{}_{\overset{\square}{w}}r_i}{\sqrt{1 - {}_{\overset{\square}{w}}h_{ii}}}$
Studentized residual	$\dfrac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$	$\dfrac{{}_{\overset{\square}{w}}e_i}{\sqrt{{}_{\overset{\square}{w}}\hat{\sigma}^2 w_i^{-1}\left(1 - {}_{\overset{\square}{w}}h_{ii}\right)}} = \dfrac{{}_{\overset{\square}{w}}r_i}{\sqrt{{}_{\overset{\square}{w}}\hat{\sigma}^2\left(1 - {}_{\overset{\square}{w}}h_{ii}\right)}}$
Studentized deletion residual*	$t_i = \dfrac{e_i}{\sqrt{\hat{\sigma}_{(i)}^2(1 - h_{ii})}}{}^{1}$	

$*$ $\hat{\sigma}_{(i)}^2$ is the estimate of σ^2 when the entire regression is run again on the n sample without the i-th case.

Appendix A. Statistical tests for testing heteroscedasticity.

For completeness, heteroscedasticity can be formally tested by some statistical tests such as the Breusch-Pagan-Godfrey test [1A] and Godfrey [2A,3A] which has been extended by Cook and Weisberg [4A]. The test statistic is asymptotically distributed as a χ^2 distribution with k degrees of freedom (where k is the number of the predictors) under the null hypothesis of homoskedasticity and of normal distribution of the residual.

Since the Breusch-Pagan-Godfrey test statistic may not be accurate for non-normal data Bickel [5A] and Koenker [6A], among others, have proposed some variants. Furthermore, it has to be noted that this test is also not robust to multicollinearity.

In the open source R language, Breusch-Pagan-Godfrey test is performed by the 'ncvTest' function available in the "car" package (https://r-forge.r-project.org/projects/car/), by the 'bptest' function available in the "lmtest" package (https://CRAN.R-project.org/package=IMTest), by the 'plmtest' function available in the "plm" package (https://cran.r-project.org/package=plm) or by the 'breusch_pagan' function available in the "skedastic" package (https://CRAN.R-project.org/package=skedastic). The Koenker's variant is implemented in the package "lmtest" (https://CRAN.R-project.org/package=IMTest) of the open-source R language.

In SAS®, Breusch–Pagan can be obtained using the Proc Model. [7A]

White's test [8A] and Cook and Weisberg's test [9A] are another statistical test for heteroscedasticity that follows a chi-squared distribution, with degrees of freedom equal to number of estimated parameters minus 1 (a constant must be included). Furthermore, Waldman [10A] showed that White's test is equivalent to the algebraically modified Godfrey and Breusch-Pagan with choice of regressors as in White's test. The White's test can be performed by the "bptest" function from the "lmtest" package of the open-source R language.

Finally, we would like to mention the Goldfeld-Quandt test in its parametric and non-parametric version [11A] even if it is usually referred in its parametric form [12A,13A].

The Goldfeld–Quandt test checks for homoscedasticity in regression analyses by dividing the dataset into two groups (for this reason the test is sometimes called a two-group test) not necessarily of equal size nor contain all the observations. The groups are specified so that the observations for which the pre-identified explanatory variable takes the lowest values are in one subset, with higher values in the other. The test statistic uses the ratio of the mean square residual errors for the regressions on the two subsets and corresponds to an F-test of equality of variances. It should be remembered that the parametric test assumes that the errors have a normal distribution and that the design matrices for the two subsets of data are both of full rank. Unfortunately, the Goldfeld–Quandt test is not very robust to specification errors. Thursby [14A] proposed a modification of the Goldfeld–Quandt test by using a variation of the Ramsey's "RESET" test [15A] in order to obtain some measure of its robustness.

In the open source R language, the Goldfeld-Quandt Test can be implemented using the 'gqtest' function of the "lmtest" package ((https://CRAN.R-project.org/package=IMTest) (parametric test only) or using the 'goldfeld_quandt' function of the "skedastic" package (https://CRAN.R-project.org/package=skedastic). (both parametric and nonparametric test).

References for Appendix A.

1A. Breusch TS, Pagan AR, A Simple Test for Heteroskedasticity and Random Coefficient Variation, Econometrica, 1979: 47 (5);1287–94, DOI:10.2307/1911963

2A. Godfrey L. Testing against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables Econometrica, 1978: 46 (6); 1293-301.

3A. Godfrey L. Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables Econometrica 1978: 46 (6); 1303-10.
4A. Cook and Weisberg (Cook RD, Weisberg S. Diagnostics for Heteroskedasticity in Regression, Biometrika 1983:70(1);1–10, DOI:10.1093/biomet/70.1.1.1.
5A. Bickel PJ. 1978, Using residuals robustly I: Test for heteroscedasticity and nonlinearity, Annals of Statistics 1978; 6, 266-291.
6A. Koenker R. A note on studentizing a test for heteroscedasticity, Journal of Econometrics 1981: 17;107-l12.)
7A. SAS/ETS® 13.2 User's Guide. Cary, NC: SAS Institute Inc. Copyright © 2014, SAS Institute Inc., Cary, NC, USA)
8A. White HA. Heteroscedasticity - Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. Econometrica, 1980 48; 817-838.
9A. Cook RD, Weisberg S. Diagnostics for Heteroscedasticity in Regression. Biometrika 1983:70; 1-10)
10A. Waldman DM. A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity. Economics Letters 1983: 13 (2–3); 197-200.
11A. Goldfeld SM, Quandt RE. Some Tests for Homoscedasticity". Journal of the American Statistical Association. 1965: 60 (310): 539–547),
12A. https://www.geeksforgeeks.org/goldfeld-quandt-test/ and
13A. https://en.wikipedia.org/wiki/Goldfeld%E2%80%93Quandt_test
14A. Thursby J. "Misspecification, Heteroscedasticity, and the Chow and Goldfeld-Quandt Tests". The Review of Economics and Statistics 1982: 64 (2); 314–321. doi*:*10.2307/1924311)
15A. Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis". Journal of the Royal Statistical Society, Series B. *31 (2): 350–371.*

Appendix B. General overview of some robust regression procedures

Robust regression procedures: a general overview.

It should be noted that interested readers should refer to at least one of the truly excellent books on this topic (Atkinson and Riani [1], Rousseeuw and Leroy [2] Maronna et al. [3,4], Huber [5], Huber and Ronchetti [6]) referred in this text.

In addition, from the preface of the first edition of the book by Maronna et al. [3] following sentence should be considered: "Robust methods have a long history that can be traced back at least to the end of the 19th century with Simon Newcomb [1B]. But its first great steps forward occurred in the 60s and the early 70s with the fundamental work of John Tukey [2B,3B] Peter Huber [4B,5B,6B] and Frank Hampel [7B,8B]. The applicability of the new robust methods proposed by these researchers was made possible by the increased speed and accessibility of computers."

It is also worth considering, always from the preface of the book from Maronna et al. [3], this other sentence: "robust methods remain largely unused and even unknown by most of the communities of applied statisticians, data analysts, and scientists that might benefit from their use. It is our hope that this book will help to rectify this unfortunate situation".

The well-known methods of robust estimation are: M estimation, S estimation, LST estimation and MM estimation. [9B].

1)-The M estimation, introduced by Huber [4B,10B] is the simplest approach both from a computational and theoretical point of view. Although it is not robust to leverage points, it is still widely used in data analysis when it can be assumed that contamination is mainly in the direction of the response. In fact, when there are outliers in the explanatory variables (leverage points), the method has no advantage over least squares.

The “M” in M-estimation stands for "maximum likelihood type".
There are several methods available to compute location and/or scale M-estimators. In principle one could use any of the general equation solving methods for such as the Newton–Raphson algorithm, but the computational method, called iterative reweighting, shows some advantages according to Maronna et al. [4, page 40].

2)-In the 1980s, several alternatives to M-estimation were proposed, such as the S-estimation. This method finds a line (plane or hyperplane) that minimizes a robust estimate of the scale (hence the method gets the S in its name) of the residuals [Rousseeuw and Leroy (1987, p. 263) [12B]. This method is highly resistant to leverage points and is robust to outliers in the response. However, this method was also found to be inefficient with some computational concerns according to Huber and Ronchetti (2009, p. 197) [10B]. Finally, Rocke [11B] showed that S-estimators can be sensitive to outliers even if the breakdown point is close to 0.5.

3)-A viable alternative is the Least Trimmed Squares (LTS) that was introduced by Rousseeuw [12B] and that is the preferred choice of Rousseeuw and Leroy [13B] and Ryan [14B] books, very useful for a pragmatic review of this topic. Least Trimmed Squares (LTS) estimation is a high breakdown value method, taking into account that the breakdown value is a measure of the proportion of contamination that an estimation method can withstand and still maintain its robustness. The performance of this method has been improved by the FAST-LTS algorithm proposed by Rousseeuw and Van Driessen [15B,16B]

4)-MM estimation, introduced by Yohai [17B,18B,19B] combines the high breakdown value estimation and the M estimation. It has the same high breakdown property as the S estimation but higher statistical efficiency.
MM-estimates have become increasingly popular and are one of the most commonly employed robust regression techniques.
The method proceeds by finding a highly robust and resistant S-estimate that minimizes an M-estimate of the scale of the residuals (the first M in the method's name). The estimated scale is then held constant while a close M-estimate of the parameters (the second M) is located.
Specifically, the MM estimation is based on the following three steps: A)-In the first step, it computes an initial consistent estimate $\hat{\beta}_0$ with a high breakdown point but possibly low normal efficiency; B)-In the second step, a robust M-estimate of the scale $\hat{\sigma}$ of the residuals is obtained based on the initial estimate; C)-Finally, in the third step, an M-estimate $\hat{\beta}$ starting at $\hat{\beta}_0$ is calculated [4].
Furthermore, the MM-estimator has the highest possible breakdown point of 0.5, and high efficiency under normality.
For the details of the algorithms, readers are referred to the manual of the statistical software used (package “RobStatTM” - https://CRAN.R-project.org/package=RobStatTM or package “Robustbase” https://CRAN.R-project.org/package=robustbase of the open source R language for the procedures shown in Maronna et al.’s books [3,4], even if they are at a high mathematical/statistical level, and to the manual of the of SAS®“PROC ROBUSTBASE” [20B].
As a final comment, the MM-estimates and the robust and efficient weighted least-square estimator (REWLSE) proposed by Gervini and Yohai [21B] have both a high breakdown point and a high efficiency and, according to a simulation study by Yu and Yao [22B], have the best overall performance among all the robust methods compared.

References for Appendix B

1B. Stigler S, Simon Newcomb, Percy Daniell and the history of robust estimation 1885–1920, Journal of the American Statistics Association, 1973, 68, 872–879.
2B. Tukey JW. A survey of sampling from contaminated distributions, in I. Olkin (ed.) Contributions to Probability and Statistics. 1960, Stanford University Press.
3B. Tukey JW. The future of data analysis, The Annals of Mathematical Statistics 1962, 33, 1–67.
4B. Huber PJ. Robust estimation of a location parameter, The Annals of Mathematical Statistics, 1964 35, 73–101.
5B. Huber PJ. A robust version of the probability ratio test, The Annals of Mathematical Statistics, 1965 36, 1753–1758.
6B. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions, in Proceedings of Fifth Berkeley Symposium of Mathematical Statistics and Probability, 1967 vol. 1, pp. 221–233. University of California Press.
7B. Hampel FR. A general definition of qualitative robustness, The Annals of Mathematical Statistics, 1971; 42, 1887–1896.
8B. Hampel FR. The influence curve and its role in robust estimation., The Annals of Statistics, 1974 69, 383–393.
9B. https://en.wikipedia.org/wiki/Robust_regression.
10B. Huber PJ, Ronchetti EM. Robust Statistics 2nd ed. 2009 Hoboken, NJ: John Wiley & Sons Inc.
11B. Rocke DM. Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension. The Annals of Statistics 1996, 24,3,1327–1345.
12B. Rousseeuw PJ. Least median of squares regression. Journal of the American Statistical Association 1984 79, 871–880.
13B. Rousseeuw PJ, Leroy AM. (2003) [1986]. Robust Regression and Outlier Detection. John Wiley & Sons Inc.
14B. Ryan, TP. (2008) (1997). Modern Regression Methods. John Wiley & Sons Inc.
15B. Rousseeuw PJ, van Driessen K. (1999), A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41, 212–223.
16B. Rousseeuw PJ, van Driessen K. (2000), An algorithm for positive-breakdown regression based on concentration steps, in W. Gaul, O. Opitz and M. Schader (eds), Data Analysis: Modeling and Practical Applications, pp. 335–346. Springer Verlag.
17B. Yohai VJ. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. Annals of Statistics 15:642–656.
18B. Yohai VJ, Stahel WA, Zamar RH. (1991). A Procedure for Robust Estimation and Inference in Linear Regression. In Directions in Robust Statistics and Diagnostics, Part 2, edited by Stahel WA, Weisberg SW. 365–374. New York: Springer-Verlag.
19B. Yohai VJ, Zamar RH. (1997). Optimal Locally Robust M-Estimates of Regression. Journal of Statistical Planning and Inference 64:309–323.
20B. SAS Institute Inc. 2016. SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc.
21B. Gervini D, Yohai VJ. A class of robust and fully efficient regression estimators. The Annals of Statistics 2002: 30:583–616.
22B. Yu C, Yao W. (2017) Robust Linear Regression: A Review and Comparison Communications in Statistics—Simulation and Computation. 46:8, 6261-82.

Appendix C. General overview of the MO robust procedure

The MO procedure is based on two stages to obtain both robust estimates of the model parameters and a valid classification of the outlying observations.

Particularly: the first stage jointly processes the response variable (Y) and the explanatory variable (X), which form together the so-called **Z** = (**Y, X**) matrix. Of course, it is also possible to have several independent variables to form a matrix **X** with a matrix **Z** = (**Y, X**) with more than two columns. Then, the original dataset is split into two preliminary subsets (bulk and outliers) by using an approach based on the robust Minimum Covariance Determinant (MCD) according to Rousseeuw and Van Driessen [1C on page 213] calculated by minimizing the determinant of subsets of size equal to (n+p+1)/2 where "p" is the number of the parameters. Data with the MCD make up the preliminary bulk subset (pb) on which are calculated the preliminary OLS estimates of the model parameters: ${}_{\mathrm{pb}}\mathrm{a}_{\mathrm{OLS}}$, ${}_{\mathrm{pb}}\mathrm{b}_{\mathrm{OLS}}$, and ${}_{\mathrm{pb}}\hat{\sigma}_{\mathrm{OLS}}$.

These preliminary OLS estimates are used to compute the externally predicted regression diagnostics of the observations that form the preliminary outlier subset; particularly, externally predicted scaled residuals in Y-space and leverage in X-space ($\tilde{\mathrm{r}}_{\mathrm{i}}$, and $\tilde{\mathrm{h}}_{\mathrm{i}}$, respectively).

Observations with externally predicted scaled residual between the specified cut-offs (2.576 and 2.576 giving a 0.99 probability of inclusion in the interval) are either false outliers identified by MCD on **Z** matrix or good leverage points. These are all added back to the 'preliminary bulk subset' to obtain the 'confirmed bulk subset'.

These observations are labelled as "typical data" if $\tilde{\mathrm{h}}_{\mathrm{i}} \leq$ 2p/(n-m+1-2p) (where m is the number of observations in the preliminary outlier subset, p is the number of parameters and n is the sample size) or "good leverage points" if $\tilde{\mathrm{h}}_{\mathrm{i}}$ >2p/(n-m+1-2p). It has to be noted that the leverage threshold of 2p/(n-m+1-2p) has been suggested by Marubini and Orenti [2C,3C].

Consistently, the remaining outliers are now forming the "confirmed outlier subset". The "confirmed bulk subset (cb)" is now used to compute the "confirmed" OLS estimates: ${}_{\mathrm{cb}}\mathrm{a}_{\mathrm{OLS}}$, ${}_{\mathrm{cb}}\mathrm{b}_{\mathrm{OLS}}$, and ${}_{\mathrm{cb}}\hat{\sigma}_{\mathrm{OLS}}$. It is noteworthy that $\tilde{\mathrm{r}}_{\mathrm{i}}$ are approximately distributed as a standard Normal variable (Salini et al. 2016) [4C] leading to justify the adopted cut-offs of -2.576 and 2.576 corresponding respectively to the 0.005 and 0.995 quantiles of the standard Normal distribution.

In the second stage, the confirmed OLS estimates start the iterative regression process of the M estimator: the weights are computed by using the biweight Tukey redescending function [5C] with constant c = 4.685 and keeping the scale parameter ${}^{\square}_{\mathrm{cb}}\hat{\sigma}_{\mathrm{OLS}}$ fixed at each iteration. The final MO estimates are so attained: $\mathrm{a}_{\mathrm{MO}}, \mathrm{b}_{\mathrm{MO}}$, and $\sqrt{\mathrm{V}}_{\mathrm{MO}}$.

Where V_{MO} is the variance covariance matric corrected according to Maronna et al. [6C, pp. 100–101).

To label the different types of observations according to the final MO estimates, a graph is provided that plots the weights (ranging from 0 to 1) of the final iteration of the MO procedure, against the natural logarithm of the square root of the robust distance (ln $_{\mathrm{Z}}$RD):

${}^{\square}_{\mathbf{z}}\mathrm{RD}_{\mathrm{i}} = \sqrt{\left(\mathbf{z}_{\mathrm{i}} - {}^{\square}_{\mathrm{pb}}\bar{\mathbf{z}}\right)' \, {}^{\square}_{\mathrm{pb}}\mathbf{S}^{-1}\left(\mathbf{z}_{\mathrm{i}} - {}^{\square}_{\mathrm{pb}}\bar{\mathbf{z}}\right)}$, where $\mathbf{z}_{\mathrm{i}} = (\mathrm{y}_{\mathrm{i}}, \mathrm{x}_{\mathrm{i}})$ and ${}^{\square}_{\mathrm{pb}}\bar{\mathbf{z}}$, ${}^{\square}_{\mathrm{pb}}\mathbf{S}$ are the corresponding estimates of the mean vector and the covariance matrix computed on the preliminary bulk subset identified in the first stage. The Robust Distance indicates how far any observation is from the center of the ellipsoid forming the bulk, and the cut-off of such distance is shown in the plot by a vertical line drawn at the natural logarithm of the squared root of the 0.95 quantile of the χ2 distribution with degrees of freedom equal to the number of parameters. The Robust Distances together with the final MO weights allow for definitive labelling of the observations (see Figure 5A and 5B together with the pertinent comments). Note that the observations forming the "good leverage" set of points were in the first step labelled as outliers by the MCD estimator, but are subsequently not considered as outliers and thus are not down-weighted in the final MO iteration.

To decide whether it makes sense to eliminate all or some of the identified outliers from the original dataset it is of fundamental importance to carefully examine the observations forming the final outlier subset and consider: (i) their robust distance; (ii) the data generation process;

and (iii) their scientific/biological plausibility on the ground of subject matter knowledge. As a result, a reduced dataset can be obtained.

As a further comment, it is useful to say that a widely used measure of remoteness of observations from the centroid of this space is the Mahalanobis Square Distance [7C] but it is substantially influenced by the presence of outliers [8C]. Therefore, the MO procedure used the Robust Distances proposed by Rousseeuw and van Zomeren [9C] as an alternative.

Finally, MO procedure can be performed by an R code available on request from the corresponding author.

References for Appendix C.

1C. Rousseeuw PJ, van Driessen K. (2000), An algorithm for positive-breakdown regression based on concentration steps, in Gaul W, Opitz O, Schader M. (eds), Data Analysis: Modeling and Practical Applications, pp. 335–346. Springer Verlag.

2C. Orenti A, Marano G, Boracchi P, Marubini E. Pinpointing outliers in experimental data: the Hat matrix in Anova for fixed and mixed effects models IJPH – 2012;9:e86631-13.

3C. Orenti A, Marubini E. Robust regression analysis: a useful two stage procedure, Communications in Statistics - Simulation and Computation, 2021;50:16-37, doi: 10.1080/03610918.2018.1547400

4C. Salini S, Cerioli A, Laurini F, Riani M. (2016). Reliable Robust Regression Diagnostics. International Statistical Review, 84(1), 99–127.

5C. Beaton AE. Tukey JW. (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 16:147–85.

6C. Maronna, R. A., D. R. Martin, and V. J. Yohai. 2006. Robust statistics: Theory and methods. Chichester, UK: John Wiley and Sons.

7C. Mahalanobis PC. (1936) On the generalised distance in statistics, in Proceedings of the National Institute of Sciences of India, 2:49–55.

8C. Li X, Deng S, Li L, Jiang Y. (2019) Outlier Detection Based on Robust Mahalanobis Distance and Its Application Open Journal of Statistics 9:15-26.

9C. Rousseeuw PJ, van Zomeren BC. (1990) Unmasking multivariate outliers and leverage points. J Am Stat Ass 85:633–639.