Gastroenterology and Hepatology From Bed to Bench.
2023 **RIGLD**, Research Institute for Gastroenterology and Liver Diseases

ORIGINAL ARTICLE

p-ISSN: 2008-2258 e-ISSN: 2008-4234

A Bayesian approach to correct the under-count of cancer registry statistics before population-based cancer registry program

Hadis Barati[1], Mohamad Amin Pourhoseingholi[2], Gholamreza Roshandel[3], Seyed Saeed Hashemi Nazari[1], Esmaeil Fattahi[4]

[1]Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3]Golestan Research Center of Gastroenterology and Hepatology, Golestan University of Medical Sciences, Gorgan, Iran
[4]Department of Health Education and Promotion, School of Health, Guilan University of Medical Sciences, Rasht, Iran

ABSTRACT

Aim: This study aims to correct undercounts in cancer data before initiating a population-based cancer registry program, employing an innovative Bayesian methodology.

Background: Underestimation is a widespread issue in cancer registries within developing countries.

Methods: This secondary study utilized cancer registry data. We employed the Bayesian approach to correct undercounting in cancer data from 2005 to 2010, using the ratio of pathology to population-based data in the Golestan province as the initial value.

Results: The results of this study showed that the lowest percentage of undercounting belonged to Khorasan Razavi province with an average of 21% and the highest percentage belonged to Sistan and Baluchestan province with an average of 38%.

The average age-standardized incidence rate (ASR) for all provinces of the country except Golestan province was equal to 105.72 (Confidence interval (CI) 95% 105.35-106.09) per 100,000 and after Bayesian correction was 137.17 (CI 95% 136.74-137.60) per 100,000. In 2010 the amount of ASR before Bayesian correction was 100.28 (CI 95% 124.39-127.09) per 100,000 for women and 136.49 (CI 95% 171.20-174.38) per 100,000 for men. Also, after implementing the Bayesian correction, ASR increased to 125.74 per 100,000 for women and 172.79 per 100,000 for men.

Conclusion: The study demonstrates the effectiveness of the Bayesian approach in correcting undercounting in cancer registries. By utilizing the Bayesian method, the average ASR after Bayesian correction with a 29.74 percent change was 137.17 per 100,000. These corrected estimates provide more accurate information on cancer burden and can contribute to improved public health programs and policy evaluation. Furthermore, this research emphasizes the suitability of the Bayesian method for addressing underestimation in cancer registries. It also underscores its pivotal role in shaping the trajectory of future investigations in this field.

Keywords: Cancer, Registry, Bayesian method, Underestimation, Iran.

(Please cite as: **Barati H, Pourhoseingholi MA, Roshandel G, Hashemi Nazari SS, Fattahi E. A Bayesian approach to correct the under-count of cancer registry statistics before population-based cancer registry program. Gastroenterol Hepatol Bed Bench 2023;16(4):421-431. https://doi.org/10.22037/ghfbb.v16i4.2843).**

Introduction

From rural villages to busy cities, the widespread impact of cancer, a multifaceted and extensive illness, highlights the pressing requirement for precise information and creative strategies to tackle its effects. (1, 2). In the field of cancer research, accurate and reliable data play a crucial role in understanding the disease, evaluating treatment options, and developing effective interventions (3, 4). Undercounting, which refers to the failure to capture all instances of cancer within a population, has the potential to introduce bias and hinder advancements in cancer research.

Received: 09 July 2023 *Accepted*: 02 September 2023
Reprint or Correspondence: **Seyed Saeed Hashemi Nazari,** *Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Tehran, Iran.*
E-mail: saeedh_1999@yahoo.com
ORCID ID: 0000-0002-0883-3408

Copyright © 2023, Gastroenterology and Hepatology From Bed to Bench (GHFBB). This is an open-access article, distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/) which permits others to copy and redistribute the material just in noncommercial usages, provided the original work is properly cited.

Underestimate can occur due to reasons such as misclassification, incomplete data collection, or a lack of awareness among healthcare providers. Consequently, the actual number of cancer cases may exceed the reported figures, leading to an inaccurate depiction of the disease burden (5). Addressing underestimate in cancer data is vital for precise epidemiological analyses, effective public health planning, and appropriate resource allocation. Failing to correct undercounting may cause researchers and policymakers to underestimate the true prevalence of cancer, resulting in inadequate support and interventions (6). Additionally, undercounting can compromise the accuracy of cancer registries, which are indispensable for tracking trends, evaluating interventions, and monitoring long-term outcomes (7).

Various methods are commonly used to correct Underestimate in cancer data. These methods include the capture-recapture method (8), machine learning (9), and data linkage and integration (10). Despite their effectiveness, these methods have certain weaknesses that warrant consideration. Machine learning approaches heavily rely on data quality and completeness, making inaccurate or incomplete data liable to produce biased or unreliable results (9). The capture-recapture method assumes a closed population, which may not hold true in real-world scenarios. Errors or incomplete data during the capture or recapture stages significantly impact the validity of the method and introduce bias. Moreover, the method requires a sufficiently large sample size to ensure reliable outcomes, which can be challenging when dealing with rare populations or small sample sizes (11, 12). Similarly, the probabilistic linkage method, another commonly used approach, possesses its own limitations. The effectiveness of probabilistic linkage heavily depends on the selection of matching variables, and utilizing inappropriate variables can lead to incorrect matches or missed links (13).

On the other hand, the Bayesian approach provides a powerful statistical framework by incorporating prior knowledge and beliefs, enabling more accurate estimation of the true number of cancer cases (14). This approach combines observed data with prior information, offering a flexible and robust method to correct undercounting bias. The Bayesian undercount correction method offers a range of valuable attributes, such as flexibility, the ability to quantify uncertainty, adaptive modeling, and the capacity to address intricate scenarios (15-18). This article delves into the undercount correction method in cancer, with a specific focus on the Bayesian approach and its advantages. Given the highlighted benefits of the Bayesian method, the research team opted to employ this approach to correct the cancer registration data in Iran. The objective of our study is to apply the Bayesian approach to correct undercounts in pathology-based cancer registry data (PaCR) in Iran from 2005 to 2010.

Methods

Statistical analysis

In this study, we corrected the undercounting percentage of pathology-based data from 2005-2010 in Iranian provinces using a Bayesian method and data from Golestan population based cancer registry (GPCR).

Two vectors Y1 and Y2 were used to enter the data into the Bayesian model. The vector $Y_1=[Y_{11}, Y_{21}, \dots, Y_{r1}]'$ demonstrate the registered values of PaCR in the provinces of Iran (except Golestan) and $Y_2=[Y_{12}, Y_{22}, \dots, Y_{r2}]'$ demonstrate the registered values of GPCR. The subscript r is the number of covariate patterns for age and sex group combinations.

For two vectors of Y1 and Y2, the Poisson distribution was considered as follows: $Y_1 \sim \text{Poisson}(P_i \mu_{i1})$ and $Y_2 \sim \text{Poisson}(P_i \mu_{i2})$ in which μ_i is the observed rate of cancer incidence for the covariate pattern. Let θ be equal to: 1 minus the ratio of pathology-based to population-based data in Golestan province; assuming the non-informative prior distribution of beta, ie θ~Beta(a,b). If the actual rate of cancer incidence for each vector is supposed to be as λ_i, the relation between actual rate and observed rate can be written in the following form; $\mu_{i1} = \lambda_{i1}(1-\theta)$ and $\mu_{i2} = \lambda_{i2} + \lambda_{i1}\theta$.

Since θ is an unknown parameter, we applied the latent variable approach proposed by Paulino et al. (19, 20), Liu et al. (21) and Stamey et al. (22); defining $U_i \mid \lambda_1, \lambda_2, \theta, Y_1, Y_2 \sim \text{Binomial}(Y_{i2}, P_i)$, where, $P_i = \frac{\lambda_{i1}\theta}{\lambda_{i1}\theta+\lambda_{i2}}$ the number of cancer cases in all provinces is re-estimated and the posterior distribution is obtained as follows.

$$\theta \mid \lambda_1, \lambda_2, U_i, Y_1, Y_2 \sim \text{Beta}(\sum_i U_i + a, \sum_i Y_{i1} + b)$$

To achieve the appropriate prior, we used the ratio of pathology to the population base in the cancer registry

data of Golestan province for each age group and each year. So, Bayesian correction was performed for every age category for achieving θ's posterior (according to the ASR definition; 0-14, 15-49, 50-69, and +70 years age groups) and for each province, separately. Afterward, by estimating the posteriors of the undercounting percentage and using appropriate proportions, the corrected "cancer cases" and "ASR" were reported from 2005 to 2010. All analyses of the present research were carried out using R software, version 4.1.2.

Data sources

The Bayesian method is highly suitable for dealing with intricate situations characterized by numerous factors contributing to undercounting, diverse populations, and varying levels of data quality. Taking into account these characteristics and the data obtained in this study, we employed the Bayesian method to correct undercounts.

The current study is a secondary study (secondary data analysis) conducted using data from the cancer registry system in Iran. To correct the undercounts of cancer data using the Bayesian method, we used PaCR data in Iran from 2005 to 2010. It is important to note that this data predates the establishment of the population-based cancer registration program (PBCR). To use the Bayesian method, a prior is required. Fortunately, cancer registration in the Golestan province of Iran has been population-based since 2005 (23), providing the ratio of pathology-based to population-based data. Therefore, we use the cancer registry data in Golestan province as a prior (ratio of pathology-based to population-based) to correct for the underestimation of the cancer registry in Iran.

The study was approved by the Medical Ethics Committee of Shahid Beheshti University of Medical Sciences (IR.SBMU.PHNS.REC.1399.132).

A cancer registration program based on pathology (reports collected from pathology centers, both governmental and non-governmental) has been implemented in the country since 1986 and its report has been published (24, 25).

Cancer incidence data from 2005 to 2010 were extracted for this study. Variables such as year of diagnosis, patient age, and sex were also collected. Annual population information is also provided by the Iranian Statistics Office.

Since the comparison of Simple Crude rates, which are the sum of cancers in the whole population, regardless of age groups, creates erroneous images, the age-standardized rates (ASRs) using the direct standardization method were calculated by gender for all provinces. ASR was calculated for 4 age groups 0-14, 15-49, 50-69, and above 70 per, 100000 ($ASR = \sum_i (W_i \times a_i)$) (26). The basis of this method is to select a standard population and calculate the desired outcome of this population by using age-specific rates of the community. The most common standard population used is World Standard Population. In this study, the WHO standard population in 2000 was used to calculate ASR in Iran.

Cancer registration program based on population of Golestan province

The GPCR was designed and launched in 2005 as a joint research project between the Gastroenterology and Liver Research Center of Golestan University of Medical Sciences and the Gastroenterology and Liver Diseases Research Institute of Tehran (23, 27).

The data of GPCR as the only source of cancer from Iran has been published in the book "Cancer Statistics in Five Continents" and is currently operating as a high-quality and active population-based cancer registration center (28).

The main sources of data in the GPCR were pathology centers, hospitals, and death cancer registration data. Other data sources such as cancer clinics, radiotherapy centers, and addiction drug control units. Were also considered as potential sources for collecting data from cancer patients (23).

In the GPCR, while performing the usual quality control criteria of the data, the following quality criteria were also controlled: Percentage of cases with morphological verified diagnosis (MV%), Percentage of cases for which the only information came from a death certificate (DCO%), percentage of cases with unknown age and also other quality control criteria, such as temporal variations and differences between different populations, were also periodically monitored (29).

In Golestan province, cancer registry data include cases based on pathology and population. The ratio of these two is also known from 2005 to 2010. We don't have this ratio for other provinces of the country (28 provinces) and only have data based on pathology.

Table 1. Bayesian estimation of underreporting percentage in Iranian provinces, 2005–2010.

Provinces	Estimated underreporting rate					
	2005	2006	2007	2008	2009	2010
Tehran	0.35	0.37	0.24	0.17	0.18	0.18
Qom	0.32	0.32	0.24	0.21	0.30	0.27
Qazvin	0.29	0.28	0.22	0.25	0.23	0.23
Mazandaran	0.34	0.29	0.24	0.20	0.21	0.22
Isfahan	0.28	0.27	0.20	0.22	0.18	0.18
Azerbaijan, East	0.44	0.48	0.19	0.19	0.20	0.24
Khorasan, Razavi	0.27	0.24	0.20	0.19	0.19	0.17
Khorasan, South	0.35	0.33	0.28	0.29	0.28	0.28
Khuzestan	0.30	0.29	0.21	0.16	0.16	0.18
Fars	0.29	0.29	0.19	0.20	0.19	0.17
Kerman	0.36	0.30	0.25	0.24	0.18	0.19
Markazi	0.35	0.34	0.26	0.28	0.18	0.19
Gilan	0.32	0.34	0.23	0.24	0.24	0.17
Azerbaijan, West	0.26	0.28	0.21	0.25	0.22	0.21
Sistan and Baluchestan	0.47	0.47	0.34	0.37	0.33	0.29
Hormozgan	0.40	0.38	0.30	0.31	0.33	0.29
Zanjan	0.32	0.34	0.24	0.27	0.27	0.26
Kermanshah	0.29	0.32	0.21	0.23	0.22	0.23
Kurdistan	0.29	0.31	0.22	0.19	0.20	0.20
Hamedan	0.29	0.31	0.23	0.22	0.22	0.21
Chahar Mahaal and Bakhtiari	0.28	0.29	0.21	0.23	0.24	0.23
Lorestan	0.32	0.34	0.16	0.21	0.21	0.19
Ilam	0.41	0.33	0.22	0.26	0.26	0.25
Kohgiluyeh and Boyer-Ahmad	0.32	0.29	0.21	0.25	0.20	0.21
Semnan	0.40	0.37	0.22	0.17	0.21	0.20
Ardabil	0.34	0.28	0.26	0.22	0.22	0.22
Yazd	0.32	0.36	0.21	0.18	0.17	0.16
Bushehr	0.32	0.29	0.23	0.25	0.25	0.25
Khorasan, North	0.40	0.39	0.23	0.27	0.25	0.25

Therefore, we used the ratio of Golestan province (based on year and age subgroups) as the initial value in the Bayesian model.

One minus the ratio of pathology-based to population-based data (for each year and age category) in Golestan province ($\theta = 1 - \frac{\text{the morphological diagnosis cases}}{\text{the total number of diagnosed cases from all sources}}$) (i.e. a prior) was used for the correction of cancer estimates in each province for different years and age categories (refer to the statistical analysis).

The International Classification of Diseases, 10th edition (ICD10) is used to classify cancer cases. The criteria for the inclusion of cancer cases in this study was only malignant tumors (cases with behavior code 3). In other words, cases with behavior codes zero, one, and two are not included in the analysis.

Results

During 6 years, the total number of cancers was 361,203, which was 11,311 for the GPCR and 349,892 in other provinces (PaCR). The average ASR for all provinces of the country except Golestan province was equal to 105.72 (Confidence interval (CI) 95% 105.35-106.09) per 100,000 persons. The average ASR for Golestan province in the population-based cancer registry and pathology-based cancer registry was equal to 174.42 (CI 95% 171.05-177.79) and 126.78 (CI 95% 123.92-129.64) per 100,000 persons, respectively.

The ratio of pathology-based to the population-based cancer registry in Golestan province was 0.77. One minus this ratio was 0.23. In the age groups of 0-14 years, 15-49 years, 50-69 years, and +70 years, this ratio was 0.93, 0.76, 0.75, and 0.67 respectively. We didn't have these ratios in other provinces. After Bayesian estimation, the underestimation ratio in these provinces was 0.26 on average.

The lowest percentage of undercounting belonged to Khorasan Razavi province with an average of 21% and the highest percentage belonged to Sistan and Baluchestan province with an average of 38% (Table 1). Furthermore, with increasing the age of people, the percentage of

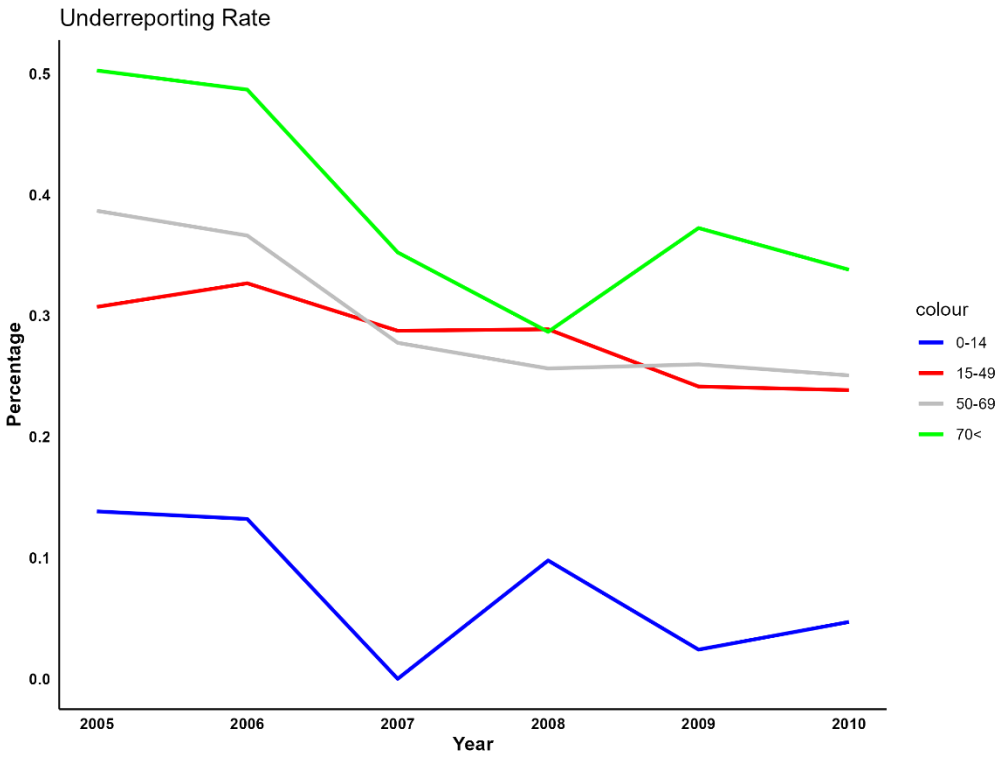

Figure 1. Bayesian estimation of underreporting percentage in any age categories, 2005–2010. The graph displays the percentage of undercounts in different age groups, with the blue, red, grey, and orange lines indicating the undercount percentages for age groups under 14 years, 15 to 49 years, 50 to 69 years, and 70 years and older, respectively. The lowest percentage of underreporting was observed in the age group from 0 to 14 years and, in contrast, the highest percentage of underreporting was observed in the age group 70 years and older.

undercounting increased. Undercounting percentages in each age category are shown in the Figure 1.

The average ASR after Bayesian correction was 137.17 (CI 95% 136.74-137.60) per 100,000. During 6 years, the highest ASR was observed in Khorasan Razavi (174.23, CI 95% 172.51-175.95), Khuzestan (171.45, CI 95% 169.24-173.66), Yazd (163.26, CI 95% 159.40-167.12), and Isfahan (162.03, CI 95% 160.31-163.75) respectively. Sistan and Baluchistan (73.08, CI 95% 70.94-75.22) and Hormozgan (86.64, CI 95% 83.92-89.36) had the lowest ASR (Table 2).

Before Bayesian correction, the ASR value for all cancers in 2005 was 69.24 per 100,000 (CI 95% 68.16-70.32) for women (16193 cases) and 89.52 per 100,000 (CI 95% 88.03-90.47) for men (20838 cases). However, after Bayesian correction, the ASR increased to 95.70 per 100,000 (CI 95% 94.45-96.95) for women (21,936 cases) and 125.18 per 100,000 (CI 95% 123.77-126.59) for men (28,826 cases). In 2010, the ASR before Bayesian correction was 100.28 per 100,000 (CI 95% 99.06-101.50) for women (32592 cases) and 136.49 per 100,000 (CI 95% 135.06-137.92) for men (40,868 cases). After implementing the Bayesian correction, the ASR increased to 125.74 per 100,000 (CI 95% 124.39-127.09) for women (40156 cases) and 172.79 per 100,000 (CI 95% 171.20-174.38) for men (50,866 cases). (See Tables 3 and 4 for details). After Bayesian correction, the total number of cancer cases was 446,158 (27.5% increase) (Table 5).

Discussion

The results of this study, utilizing the Bayesian method to correct the cancer registry data, demonstrated that the province with the highest percentage of underreporting was Sistan and Baluchestan. Notably, the underreporting rate in this province decreased from 47% in 2005 to 29% in 2010. Similarly, Khorasan Razavi province had the lowest percentage, which decreased from 27% in 2005 to 17% in 2010. This positive trend of reduction in underreporting was observed across all provinces of Iran, possibly indicating an improvement in the cancer registry reporting system, particularly from 2007 onwards. It is

Table 2. Age standardized rate of cancer incidence before and after Bayesian correction in Iranian provinces, 2005–2010.

Provinces	Before Bayesian correction						After Bayesian correction					
	2005	2006	2007	2008	2009	2010	2005	2006	2007	2008	2009	2010
Tehran	68.47	67.46	99.25	175.73	158.18	152.60	97.77	96.77	129.75	210.87	194.28	186.46
Qom	84.25	91.74	101.45	130.34	67.83	83.35	116.65	125.86	132.58	163.00	94.12	110.97
Qazvin	97.09	99.51	110.97	116.41	112.63	106.52	130.88	134.19	143.10	147.74	145.10	137.04
Mazandaran	77.88	97.06	95.35	132.54	119.94	106.93	108.97	131.87	125.36	164.74	152.17	137.07
Isfahan	103.54	113.88	126.94	128.58	148.72	143.96	138.61	150.91	160.47	160.84	184.05	177.32
Azerbaijan, East	43.42	37.43	138.35	151.49	139.42	99.24	65.93	58.23	172.86	185.37	174.21	128.56
Khorasan, Razavi	105.55	137.84	131.79	155.51	145.68	156.29	140.69	177.88	165.73	189.76	180.82	190.49
Khorasan, South	69.82	76.13	75.74	76.12	77.86	78.57	98.99	106.69	102.85	102.18	105.16	105.30
Khuzestan	87.25	98.40	117.43	194.33	176.59	148.02	119.53	133.11	149.76	230.41	213.97	181.93
Fars	94.80	94.51	140.06	136.43	138.47	158.04	128.21	128.45	174.57	169.14	172.99	192.43
Kerman	65.81	93.29	91.60	112.65	154.81	136.76	94.13	127.17	120.98	143.37	190.60	169.72
Markazi	68.28	77.08	82.20	86.89	151.02	136.41	97.20	108.17	110.31	114.89	186.41	168.93
Gilan	85.83	75.72	102.12	108.33	100.01	156.49	117.83	106.61	133.09	138.63	130.82	190.82
Azerbaijan, West	117.17	109.89	122.82	102.10	121.61	124.48	153.90	146.87	155.93	131.95	155.26	156.70
Sistan and Baluchestan	35.92	34.91	52.89	52.03	59.22	73.17	55.05	53.66	74.99	73.94	82.09	98.73
Hormozgan	49.54	56.72	66.67	68.69	58.81	70.93	73.54	82.32	92.04	93.36	82.69	95.86
Zanjan	81.62	80.41	96.41	83.51	80.25	81.12	113.42	112.39	126.47	110.48	108.35	107.89
Kermanshah	101.76	87.92	122.49	129.57	116.94	109.84	136.74	121.33	155.80	162.11	150.01	140.68
Kurdistan	103.08	102.53	111.44	158.45	137.11	130.37	138.53	138.63	143.47	192.88	172.19	163.04
Hamedan	86.54	89.18	100.23	127.33	117.27	116.33	118.26	122.22	130.79	159.82	150.15	147.24
Chahar Mahaal and Bakhtiari	92.96	153.14	140.96	110.02	101.82	101.43	125.74	206.15	180.08	140.58	132.61	130.48
Lorestan	83.21	86.47	174.44	144.80	121.21	137.25	115.38	119.56	211.84	178.02	154.02	170.40
Ilam	49.98	83.34	113.08	103.34	87.22	86.32	74.94	115.39	145.07	132.52	116.26	113.75
Kohgiluyeh and Boyer-Ahmad	82.11	97.11	123.37	109.08	131.57	124.14	114.77	132.02	156.63	139.66	165.24	156.43
Semnan	54.26	59.14	107.29	180.80	122.51	136.83	80.29	86.14	138.55	215.34	155.78	169.37
Ardabil	74.37	111.85	84.09	128.48	118.86	115.68	104.80	149.08	112.04	160.51	152.22	147.15
Yazd	77.88	67.54	122.27	163.27	169.88	180.39	108.58	96.06	155.22	197.55	206.61	215.57
Bushehr	79.59	93.41	105.33	99.29	97.88	87.09	110.06	126.82	136.44	128.42	129.06	114.84
Khorasan, North	49.90	59.10	109.17	96.58	94.98	94.66	73.85	85.97	141.06	125.74	125.24	123.61

worth noting that the underestimation percentages in recent years show a general convergence among the majority of provinces, with closely aligned values.

According to the results of this study, the highest percentage of underreporting was observed in the age category of 70 years and above, and the lowest percentage was observed in the age category under 14 years. Over a period of six years, the ASR for all cancers was 105.72 per 100,000 people, and after Bayesian correction, it was 137.17 per 100,000 people.

In the study conducted in 2014 on the incidence of cancer in Iran, the results showed that the number of pathology-based cancers was 76,568 cases (68.32%). Additionally, the cases of recorded death certificate (DCO) and clinical were 35569 cases (31.73%) (30). Also, our finding in 2005 and 2010 (DCO and clinical) were 33% and 22%, respectively.

The study conducted by Roshandel et al. on the PBCR in 2014 demonstrated the ASRs of all cancers were 177.44 and 141.18 per 100,000 in males and females, respectively. Also, this finding in our study was 172.79 and 125.74 per 100000 in males and females, respectively. The ASR of the current study is consistent with those of Roshandel et al (30).

The study conducted to estimate Cancer incidence in the East Azerbaijan province of Iran (20th March 2015 and 19th March 2016) results of a PBCR show The ASR for all cancers was 167.1 per 100,000 males

Table 3. Age standardized rate of cancer incidence before and after Bayesian correction in female group in Iranian provinces, 2005–2010.

Province / Female	Before Bayesian correction						After Bayesian correction					
	2005	2006	2007	2008	2009	2010	2005	2006	2007	2008	2009	2010
Tehran	62.25	62.38	89.24	151.96	135.18	131.99	88.06	88.74	116.40	182.65	165.12	160.59
Qom	74.79	76.71	86.36	117.43	57.30	75.04	102.32	104.17	112.22	146.56	78.78	99.33
Qazvin	90.37	83.96	91.29	100.03	99.42	86.25	120.87	112.10	117.47	127.16	127.06	110.58
Mazandaran	64.40	86.10	84.89	115.47	104.74	89.35	89.08	115.95	110.92	143.74	131.42	113.76
Isfahan	92.91	100.10	113.26	114.26	131.68	130.78	122.90	131.21	142.40	142.64	161.82	160.15
Azerbaijan, East	37.09	28.46	113.78	131.57	115.64	85.10	55.46	43.82	141.45	160.94	143.10	109.00
Khorasan, Razavi	94.22	116.86	115.69	134.91	123.80	135.19	124.52	149.77	145.17	164.78	152.70	164.22
Khorasan, South	67.16	71.35	68.26	72.44	70.07	67.22	94.22	98.27	91.92	96.91	93.70	89.02
Khuzestan	79.69	92.34	105.14	176.43	157.51	136.53	107.99	123.99	133.43	209.59	190.11	167.13
Fars	80.85	82.71	120.39	119.49	122.31	138.99	108.05	111.18	149.65	148.31	151.49	168.34
Kerman	57.12	87.48	83.82	99.31	138.07	113.05	80.44	118.64	110.27	126.41	169.36	139.45
Markazi	56.27	68.31	73.86	67.66	129.91	115.83	79.08	94.95	98.21	89.33	159.26	143.03
Gilan	72.32	62.39	85.14	87.92	82.44	126.27	97.92	87.11	110.60	112.46	106.62	153.33
Azerbaijan, West	88.12	85.19	98.98	82.21	93.19	100.59	114.82	113.08	125.56	106.22	118.21	126.03
Sistan and Baluchestan	32.43	32.73	48.80	48.42	53.53	69.21	48.79	49.64	68.49	68.50	73.25	93.13
Hormozgan	44.35	57.50	65.75	67.04	60.07	64.00	65.03	82.56	89.99	90.43	83.80	85.64
Zanjan	61.44	62.41	69.57	63.93	56.19	58.40	84.70	86.80	91.29	84.55	74.78	77.03
Kermanshah	86.40	77.27	110.18	113.58	102.86	98.13	115.38	106.09	139.88	142.13	130.89	124.69
Kurdistan	80.74	82.42	88.68	132.49	116.16	104.43	107.98	111.09	113.84	161.24	145.10	130.15
Hamedan	73.35	74.90	80.20	97.53	96.90	90.84	98.97	101.82	104.31	122.45	123.45	114.45
Chahar Mahaal and Bakhtiari	77.50	75.43	86.53	88.45	83.99	81.17	102.84	100.38	109.71	112.74	108.26	103.64
Lorestan	74.59	73.16	152.84	123.27	97.50	109.45	102.65	100.25	185.37	151.84	123.29	135.35
Ilam	46.56	75.61	83.64	84.10	74.19	79.89	69.08	104.69	107.10	108.09	98.28	104.88
Kohgiluyeh and Boyer-Ahmad	57.42	75.39	89.72	90.77	95.33	93.99	80.09	101.18	114.10	116.40	119.11	117.48
Semnan	52.92	52.83	88.62	155.87	110.65	119.14	78.04	76.70	114.50	186.00	140.17	146.91
Ardabil	58.27	88.77	72.88	100.88	94.48	94.13	81.62	117.92	96.87	126.18	120.45	119.48
Yazd	76.13	67.90	116.37	154.23	152.60	163.15	105.21	95.51	147.47	187.06	185.25	194.43
Bushehr	70.30	83.89	101.01	91.01	91.62	75.84	95.43	112.92	130.65	117.86	119.44	98.50
Khorasan, North	98.10	43.93	80.18	82.07	74.69	74.27	153.88	63.42	103.71	106.87	97.45	96.83

and 125.7 per 100,000 females. Also, this finding in our study was 148.11 per 100,000 males and 109.00 per 100000 females in 2010. The results of this study will now be compared to the findings of previous work (31).

In a study conducted by Jianguang et al. in Sweden, comparing cancer registration, and hospital discharge registration, and death registration from 1999 to 2008, the results showed that the Swedish Cancer Registry (CR) had no records of 10.6% of cancer cases are recorded in the Death Registry (DR). Similarly, the identification rate in the Hospital Discharge Registry (HDR) was 84.5% for concordant cancer cases, with 9.6% of cases missing. Neither source reported cancers for 3.4% of cancer cases recorded in DR. In conclusion, approximately 10% of cancer deaths had no cancer records in either CR or in HDR, and 3.4% were missing in both sources (32). The underestimation percentage was affected by tumor site and age at death. In our study, we figure out that underreporting percentage were affected by age category in patients.

Numerous studies have been conducted on the subject of correction of misclassification in the provinces of Iran. However, our study, which focuses on estimating the undercount percentage, is the first of its kind in Iran. Strengths of our study include the utilization of Bayesian analysis to estimate the

Table 4. Age standardized rate of cancer incidence before and after Bayesian correction in Male group in Iranian provinces, 2005–2010.

Province / Male	Before Bayesian correction						After Bayesian correction					
	2005	2006	2007	2008	2009	2010	2005	2006	2007	2008	2009	2010
Tehran	74.69	72.53	109.25	199.50	181.19	173.20	107.48	104.80	143.10	239.09	223.43	212.34
Qom	93.70	106.77	116.54	143.26	78.37	91.65	130.98	147.56	152.95	179.44	109.45	122.61
Qazvin	103.81	115.06	130.64	132.79	125.83	126.78	140.89	156.29	168.73	168.33	163.14	163.50
Mazandaran	91.36	108.03	105.81	149.62	135.14	124.51	128.87	147.78	139.80	185.73	172.92	160.37
Isfahan	114.18	127.65	140.61	142.90	165.76	157.14	154.33	170.61	178.55	179.04	206.27	194.48
Azerbaijan, East	49.74	46.39	162.92	171.41	163.20	113.38	76.40	72.65	204.27	209.80	205.31	148.11
Khorasan, Razavi	116.87	158.82	147.90	176.10	167.56	177.39	156.87	205.98	186.30	214.74	208.95	216.77
Khorasan, South	72.47	80.91	83.22	79.80	85.65	89.93	103.75	115.12	113.79	107.44	116.62	121.59
Khuzestan	94.81	104.47	129.72	212.22	195.66	159.51	131.08	142.23	166.08	251.23	237.82	196.73
Fars	108.76	106.31	159.73	153.38	154.62	177.09	148.36	145.72	199.48	189.98	194.48	216.53
Kerman	74.51	99.11	99.39	125.99	171.55	160.46	107.83	135.70	131.69	160.32	211.85	199.99
Markazi	80.29	85.85	90.55	106.12	172.13	156.99	115.31	121.41	122.41	140.45	213.56	194.83
Gilan	99.35	89.05	119.10	128.74	117.57	186.71	137.73	126.11	155.58	164.80	155.02	228.31
Azerbaijan, West	146.22	134.59	146.66	121.99	150.03	148.38	192.97	180.67	186.31	157.68	192.32	187.37
Sistan and Baluchestan	39.41	37.10	56.98	55.65	64.91	77.13	61.31	57.69	81.50	79.37	90.92	104.33
Hormozgan	54.73	55.94	67.59	70.34	57.55	77.87	82.04	82.09	94.08	96.30	81.57	106.09
Zanjan	101.80	98.41	123.25	103.08	104.31	103.84	142.14	137.97	161.65	136.40	141.92	138.75
Kermanshah	117.13	98.57	134.80	145.55	131.02	121.55	158.10	136.57	171.73	182.10	169.12	156.66
Kurdistan	125.43	122.64	134.21	184.40	158.07	156.31	169.08	166.17	173.10	224.52	199.28	195.94
Hamedan	99.73	103.46	120.26	157.13	137.65	141.82	137.54	142.63	157.28	197.20	176.85	180.04
Chahar Mahaal and Bakhtiari	108.42	130.84	149.70	131.60	119.66	121.68	148.64	177.69	190.63	168.42	156.95	157.32
Lorestan	91.82	99.78	196.05	166.33	144.92	165.06	128.11	138.87	238.31	204.21	184.75	205.45
Ilam	53.39	91.06	142.53	122.58	100.25	92.76	80.80	126.10	183.04	156.95	134.25	122.61
Kohgiluyeh and Boyer-Ahmad	106.80	118.83	157.01	127.39	167.81	154.29	149.46	162.85	199.16	162.91	211.37	195.38
Semnan	55.61	65.45	125.95	205.72	134.37	154.52	82.54	95.59	162.60	244.69	171.38	191.83
Ardabil	90.48	134.92	95.30	156.09	143.24	137.23	127.99	180.25	127.21	194.84	183.99	174.82
Yazd	79.62	67.19	128.18	172.31	187.16	197.62	111.95	96.61	162.97	208.04	227.96	236.70
Bushehr	88.88	102.92	109.64	107.57	104.14	98.34	124.70	140.71	142.24	138.97	138.69	131.17
Khorasan, North	62.17	74.28	138.15	111.09	115.28	115.06	92.88	108.52	178.41	144.62	153.04	150.39

undercount percentage of cancer. The Bayesian approach as a statistical method and a flexible manner in solving the problems of misclassification adjustment (33-35) and estimation of underreporting percentage, has always been remarkable due to its cost-effectiveness and high execution speed (36).

The Bayesian approach is well-suited to address complex scenarios characterized by multiple sources of undercounting, heterogeneous populations, and varying data quality. The flexibility of Bayesian models allows for the seamless integration of diverse data sources and the incorporation of covariates or auxiliary information to account for potential biases stemming from undercounting. This empowers researchers to effectively navigate intricate situations and derive more accurate estimates of the true disease burden (37, 38).

According to the characteristics of the cancer data in this study, we believe that selecting the Bayesian method to correct undercounts is highly suitable. The outcomes of this study can serve as a foundation for future research in the field of cancer.

To draw an overarching conclusion regarding the cancer burden over the past two decades, it is crucial to apply the Bayesian method to correct the pathology-based cancer data. By combining this corrected

Table 5. Total number of cancer cases before and after Bayesian correction in Iranian provinces, 2005–2010.

Provinces	Before Bayesian correction						After Bayesian correction					
	2005	2006	2007	2008	2009	2010	2005	2006	2007	2008	2009	2010
Tehran	6163	6401	9923	18441	17312	17620	8714	9119	12957	22170	21177	21452
Qom	517	601	691	945	511	636	710	821	904	1180	702	844
Qazvin	728	774	891	960	954	935	974	1036	1149	1221	1224	1202
Mazandaran	1746	2211	2292	3296	3054	2857	2423	2990	3006	4102	3865	3657
Isfahan	3440	3945	4560	4817	5752	5819	4568	5194	5758	6025	7102	7156
Azerbaijan, East	1194	1061	4009	4509	4291	3141	1801	1645	5006	5520	5351	4064
Khorasan, Razavi	3889	5235	5165	6269	6089	6770	5140	6726	6480	7654	7536	8234
Khorasan, South	330	378	388	384	403	420	469	530	526	516	546	565
Khuzestan	2083	2482	3039	5067	4801	4197	2797	3302	3841	6009	5767	5116
Fars	2706	2911	4373	4388	4671	5526	3624	3916	5438	5442	5810	6708
Kerman	1069	1576	1603	2022	2915	2676	1512	2129	2103	2572	3569	3303
Markazi	706	814	926	983	1780	1657	1003	1142	1242	1303	2202	2056
Gilan	1734	1589	2212	2429	2311	3733	2367	2232	2884	3112	3020	4554
Azerbaijan, West	2169	2085	2410	2094	2565	2716	2832	2780	3056	2708	3265	3411
Sistan and Baluchestan	405	426	678	660	798	977	607	637	931	921	1071	1289
Hormozgan	366	482	574	609	541	678	537	688	782	821	749	903
Zanjan	552	542	672	623	611	643	765	758	882	824	823	853
Kermanshah	1287	1172	1686	1841	1719	1674	1724	1612	2143	2305	2197	2138
Kurdistan	983	1003	1146	1665	1473	1456	1320	1356	1474	2030	1846	1818
Hamedan	1089	1144	1321	1738	1640	1680	1478	1563	1724	2184	2100	2128
Chahar Mahaal and Bakhtiari	505	587	679	666	628	649	675	785	862	849	815	831
Lorestan	870	943	1952	1698	1468	1775	1200	1301	2371	2091	1861	2197
Ilam	162	280	397	360	325	329	242	386	509	462	431	431
Kohgiluyeh and Boyer-Ahmad	293	383	463	430	531	534	406	514	586	552	663	669
Semnan	225	266	492	853	601	694	332	386	635	1019	763	860
Ardabil	615	939	715	1142	1088	1075	865	1251	951	1429	1391	1367
Yazd	531	467	892	1239	1337	1471	736	663	1133	1500	1625	1757
Bushehr	411	493	579	570	592	559	558	660	743	736	770	726
Khorasan, North	263	317	597	548	562	563	383	457	769	713	735	732

information with population-based cancer registration data, we can find a real trend in cancer.

As previously mentioned, one minus the ratio of pathology-based to population-based data in Golestan province was used for the correction of cancer estimates in each province. In our study, age and gender were used as covariates. It seems that this ratio might be affected by other covariates (such as rural/urban, education level, access to healthcare, pathology center, etc.) as well. The current study was unable to access these variables, and taking them into account might improve the results.

Conclusion

The highest percentage of underreporting was observed in Sistan and Baluchestan provinces, which decreased from 47% in 2005 to 29% in 2010. The lowest percentage was found in Khorasan Razavi province, which decreased from 27% in 2005 to 17% in 2010. These corrected estimates can be utilized to update cancer burden studies and to evaluate and improve public health programs. Underestimate in cancer data presents significant challenges in accurately comprehending the disease burden and developing effective interventions. Our study demonstrated that the Bayesian undercount correction method offers a promising approach to address this issue by incorporating prior knowledge, providing flexibility, and quantifying uncertainty. By employing Bayesian statistics, researchers can enhance the accuracy of cancer data, enabling more informed decision-making and efficient resource allocation.

Conflict of interests

There are no conflict of interest.

Supporting information

Supplementary File. Bayesian estimation of the underreporting percentage in Iranian provinces, from 2005 to 2010. Also, Age-standardized rate of cancer

incidence before and after Bayesian correction in Iranian provinces from 2005 to 2010 (RAR).

References

1. Mao JJ, Pillai GG, Andrade CJ, Ligibel JA, Basu P, Cohen L, et al. Integrative oncology: addressing the global challenges of cancer prevention and treatment. CA Cancer J Clin 2022;72:144-64.

2. Redondo-Sánchez D, Petrova D, Rodríguez-Barranco M, Fernández-Navarro P, Jiménez-Moleón JJ, Sánchez M-J. Socio-economic inequalities in lung cancer outcomes: an overview of systematic reviews. Cancers 2022;14:398.

3. Wanner M, Matthes KL, Korol D, Dehler S, Rohrmann S. Indicators of data quality at the cancer registry Zurich and Zug in Switzerland. Biomed Res Int 2018;2018:7656197.

4. Frech S, Muha CA, Stevens LM, Trimble EL, Brew R, Perin DP, et al. Perspectives on strengthening cancer research and control in Latin America through partnerships and diplomacy: experience of the National Cancer Institute's Center for Global Health. J Glob Oncol 2018;4:1-11.

5. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 2009;45:747-55.

6. Wei W, Zeng H, Zheng R, Zhang S, An L, Chen R, et al. Cancer registration in China and its role in cancer prevention and control. Lancet Oncol 2020;21:342-9.

7. Conway D, Purkayastha M, Chestnutt I. The changing epidemiology of oral cancer: definitions, trends, and risk factors. Br Dent J 2018;225:867-73.

8. Chao A, Tsay P, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. Stat Med 2001;20:3123-57.

9. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17.

10. Kum H-C, Krishnamurthy A, Machanavajjhala A, Reiter MK, Ahalt S. Privacy preserving interactive record linkage (PPIRL). J Am Med Inform Assoc 2014;21:212-20.

11. Stephen C. Capture-recapture methods in epidemiological studies. Infect Control Hosp Epidemiol 1996;17:262-6.

12. Bird SM, King R. Multiple systems estimation (or capture-recapture estimation) to inform public policy. Annu Rev Stat Appl 2018;5:95-118.

13. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. An overview of record linkage methods. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. 2014.

14. Li G, Shi J. Applications of Bayesian methods in wind energy conversion systems. Renew Energ 2012;43:1-8.

15. Bon JJ, Bretherton A, Buchhorn K, Cramb S, Drovandi C, Hassan C, et al. Being Bayesian in the 2020s: opportunities and challenges in the practice of modern applied Bayesian statistics. Philos Trans Royal Soc A 2023;381:20220156.

16. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to Bayesian methods in health technology assessment. Br Med J 1999;319:508-12.

17. Jack Lee J, Chu CT. Bayesian clinical trials in action. Stat Med 2012;31:2955-72.

18. Berniker M, Kording K. Bayesian approaches to sensory integration for motor control. Wiley Interdiscip Rev Cogn Sci 2011;2:419-28.

19. Paulino CD, Soares P, Neuhaus J. Binomial regression with misclassification. Biometrics 2003;59:670-5.

20. Paulino C, Silva G, Alberto Achcar J. Bayesian analysis of correlated misclassified binary data. CMStatistics 2005;49:1120-31.

21. Liu Y, Johnson WO, Gold EB, Lasley BL. Bayesian analysis of risk factors for anovulation. Stat Med 2004;23:1901-19.

22. Stamey JD, Young DM, Seaman JW, Jr. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. Stat Med 2008;27:2440-52.

23. Roshandel G, Sadjadi A, Aarabi M, Keshtkar A, Sedaghat S, Nouraie S, et al. Cancer incidence in Golestan Province: report of an ongoing population-based cancer registry in Iran between 2004 and 2008. Arch Iran Med 2012;15:0-.

24. Raeisi A, Janbabaei G, Malekzadeh R. Iranian Annual of National Cancer Registration Report, Islamic Republic Iran, Ministry of Health and Medical Education, Health and Treatment Deputy. Center for Disease Control and Prevention, Non communicable Disease Unit, Cancer office. 2008;2009:2019.

25. Etemadi A, Sadjadi A, Semnani S, Nouraie SM, Khademi H, Bahadori M. Cancer registry in Iran: a brief overview. Arch Iran Med 2008;11:577-80.

26. Ahmad O, Boschi-Pinto C, Lopez A, Murray C, Lozano R, Inoue M. Age standardization of rates: a new WHO standard. Geneva: World Health Organization 2001;9:1-14.

27. Roshandel G, Semnani S, Fazel A, Honarvar M, Taziki M, Sedaghat S, et al. Building cancer registries in a lower resource setting: the 10-year experience of Golestan, Northern Iran. Cancer Epidemiol 2018;52:128-33.

28. Forman D, Bray F, Brewster D, Gombe Mbalawa C, Kohler B, Piñeros M. Cancer incidence in five continents, volume X. IARC scientific publication No. 164. Lyon, France: International Agency for Research on Cancer. 2014.

29. Roshandel G, Semnani S, Fazel A, Bray F, Malekzadeh R, editors. Cancer epidemiology in Golestan, Iran: 10-year results of Golestan population-based cancer registry (2004-2013). Gorgan: Peyk Rehyan; 2017.

30. Roshandel G, Ghanbari-Motlagh A, Partovipour E, Salavati F, Hasanpour-Heidari S, Mohammadi G, et al. Cancer incidence in Iran in 2014: Results of the Iranian National Population-based Cancer Registry. Cancer Epidemiol 2019;61:50-8.

31. Somi M, Dolatkhah R, Sepahi S, Belalzadeh M, Sharbafi J, Abdollahi L, et al. Cancer incidence in the East Azerbaijan province of Iran in 2015–2016: results of a population-based cancer registry. BMC Public Health 2018;18:1-13.

32. Ji J, Sundquist K, Sundquist J, Hemminki K. Comparability of cancer identification among Death Registry, Cancer Registry and Hospital Discharge Registry. Int J Cancer 2012;131:2085-93.

33. Hajizadeh N, Baghestani AR, Pourhoseingholi MA, Ashtari S, Fazeli Z, Vahedi M, et al. Trend of hepatocellular carcinoma incidence after Bayesian correction for misclassified data in Iranian provinces. World J Hepatol 2017;9:704-10.

34. Hajizadeh N, Pourhoseingholi MA, Baghestani AR, Abadi A, Zali MR. Bayesian adjustment for over-estimation and under-estimation of gastric cancer incidence across Iranian provinces. World J Gastrointest Oncol 2017;9:87.

35. Pourhoseingholi MA, Faghihzadeh S, Hajizadeh E, Abadi A, Zali MR. Bayesian estimation of colorectal cancer mortality in the presence of misclassification in Iran. Asian Pac J Cancer Prev 2009;10:691-4.

36. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics 2006;24:1-19.

37. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. Biom J 2021;63:915-47.

38. Ashby D, Smith AF. Evidence-based medicine as Bayesian decision-making. Stat Med 2000;19:3291-305.