Journal of Clinical Epidemiology 114 (2019) 141−149

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Using health insurance reimbursement data to identify incident cancer cases

Chao Shi[a,1], Mengfei Liu[a,1], Zhen Liu[a], Chuanhai Guo[a], Fenglei Li[b], Ruiping Xu[c], Fangfang Liu[a], Ying Liu[a], Jingjing Li[a], Hong Cai[a], Zhonghu He[a,*], Yang Ke[a,*]

[a]*Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Laboratory of Genetics, Peking University Cancer Hospital & Institute, #52 Fucheng Road, Beijing, People's Republic of China*

[b]*Hua County People's Hospital, Wenming Road, Hua County, Henan Province, People's Republic of China*

[c]*Anyang Cancer Hospital, #1 Hengbin North Road, Anyang City, Henan Province, People's Republic of China*

Accepted 12 June 2019; Published online 18 June 2019

Abstract

Objectives: The objective of this study was to establish an optimal population-level follow-up strategy for identifying incident cancers using health insurance reimbursement data in rural China.

Study Design and Setting: We compared active follow-up and passive linkage with claims data for identification of incident cancer cases. Claims data were derived from the New Rural Cooperative Medical Scheme (NCMS). Follow-up data from subject enrollment to December 31, 2016, regarding 33,948 subjects in a large-scale randomized controlled trial were used in this study.

Results: The overall sensitivity of passive linkage with NCMS claims data was significantly higher than that of active follow-up (95.6% vs. 54.9%, $P < 0.001$). Of 12 cases missed by the NCMS data set, seven were treated on an outpatient basis and there were therefore no records in the NCMS system, and five were diagnosed at primary (township-level) health facilities and excluded from the quality control process. Of the 123 cases missed by active follow-up, 54 were reported as negative, 69 were reported as positive but had inaccurate information regarding the site of cancer, or exceeded the 6-month limitation from the date of diagnosis.

Conclusion: Passive linkage with NCMS claims data is an efficient approach for identifying incident cancers in areas without cancer registries in rural China. © 2019 Elsevier Inc. All rights reserved.

Keywords: New rural cooperative medical scheme; Linkage; Active follow-up; Cancer case ascertainment; ESECC trial; China

1. Introduction

For diseases with low incidence such as cancer, steady and efficient long-term follow-up is essential for identifying a sufficient number of cases to test research hypotheses in prospective studies [1]. Two methods have been used widely in cancer-related prospective studies for identification of new cancer diagnoses including active follow-up and passive linkage with a cancer registry.

Active follow-up through direct interaction with subjects has the advantages in flexibility and comprehensiveness of follow-up data. However, this approach is challenging and costly, especially when follow-up must be conducted for decades in a large population with low levels of education [1].

Population-based cancer registries (PBCRs) are usually considered to be an ideal data source of incident cancer diagnoses [2,3]. However, in contrast to a few developed countries (Denmark, Switzerland, and Finland) where national level cancer registries have been established for many years, registry coverage with high-quality data remains well below 10% in Africa, Asia, and Latin America [2,4,5]. In addition, most registry systems have a time delay, which may have limited its application in long-term prospective studies [3,6].

Conflicts of interest: There are no conflicts of interest with regard to publication of this study.

Funding: This work was supported by the Charity Project of the National Ministry of Health (grant number 201202014); the Natural Science Foundation of China, China (grant numbers 81473033, 81773501); the National Key R&D Program of China, China (grant number 2016YFC0901404); the Science Foundation of Peking University Cancer Hospital, China (grant number 2017-4); and the Open Project funded by Key Laboratory of Carcinogenesis and Translational Research, Ministry of Education/Beijing (grant number 2017-10).

[1] These authors contributed equally to this paper.

* Corresponding author: Tel.: +86-10-88196762; fax +86-10-88196735.

E-mail addresses: zhonghuhe@foxmail.com (Z. He); keyang@bjmu.edu.cn (Y. Ke).

https://doi.org/10.1016/j.jclinepi.2019.06.009

0895-4356/© 2019 Elsevier Inc. All rights reserved.

What is new?

Key findings

- For the first time we evaluated the performance of passive linkage with New Rural Cooperative Medical Scheme (NCMS) claims data for identifying incident cancers, as compared to active door-to-door interviews in a large-scale population-based prospective study in rural China.
- The overall sensitivity of passive linkage with NCMS claims data was 95.6% (95% CI: 92.4% −97.7%), which was significantly higher than that of active follow-up (54.9%, 95% CI: 48.8% −60.9%, $P < 0.001$).

What this study adds to what is known?

- Follow-up would be greatly simplified with high quality by using standardized data from health insurance systems, especially in regions without population level cancer registries.

What are the implications, and what should change now?

- In populations without specific cancer registries, linkage with a government-run health insurance system was an efficient means for tracking the occurrence of cancer in prospective cohorts. This strategy could also be used to establish population-based cancer registries in less developed areas.

Claims data from a health insurance system have proved to be an ideal substitute for cancer registry in identifying incident cancer cases in developed countries [7−19]. However, few studies have explored the value of health insurance reimbursement data in less developed areas, where availability of PBCRs was limited, leaving the active door-to-door interview as the only choice. The door-to-door interview had thus been adopted over long periods in many prospective studies.

The New Rural Cooperative Medical Scheme (NCMS) is a government-run health insurance program in rural China with a coverage of nearly 100%. NCMS claims data were directly recorded and uploaded in a real-time manner by health professionals in the health facilities in which the insured inpatients were diagnosed and treated. NCMS claims data thus have the potential for accurate population-level identification of cancer cases with a very short time delay. If the value of NCMS claims data in monitoring the occurrence of cancer can be demonstrated, it would benefit more than 20 ongoing large-scale prospective cohort studies in China, which have been funded by the National Science and Technology Major Project in the “13th Five-Year Plan” [20].

In this study, the performance of annual door-to-door interviews (active follow-up) and direct linkage with the NCMS claims data (passive follow-up) for identification of incident cancer cases were compared in a large-scale population-based randomized controlled trial conducted in a high-risk area of esophageal cancer in northern China. The aim of the present study was to establish an optimal population-level follow-up strategy in less developed areas in China without PBCRs.

2. Materials and methods

2.1. Study population

In January 2012, the Endoscopic Screening for Esophageal Cancer in China (ESECC) randomized controlled trial (clinical trial: NCT01688908) was initiated in Hua County, Henan Province, China, to evaluate the efficacy and cost-effectiveness of endoscopic screening for esophageal cancer. A detailed description of the original design of the ESECC has been previously published [21]. Briefly, 668 target villages of Hua County were randomly selected and allocated into the screening arm or control arm at a ratio of 1:1 (334 villages in each arm). Residents between 45 and 69 years of age who self-report no history of cancer or endoscopic examination within 5 years in the screening arm were assigned to undergo standard endoscopic examination and biopsy with iodine staining. No screening was undertaken in the control arm. All cohort participants were followed up through annual door-to-door interviews, during which vital events, including onset of cancer and death from all causes, were recorded. Data regarding cancer occurrence and death were also collected from the NCMS of Hua County and from the Death Registry of National Center for Disease Control and Prevention, respectively.

Enrollment in the ESECC trial was completed by September 2016, and 33,948 participants were included. The first annual follow-up was initiated on December 1, 2016 and was finalized at the end of April 2017. This evaluation was based on the first annual follow-up data of the ESECC trial.

2.2. Definition of reportable cancer cases

Reportable cancer cases included all newly diagnosed primary cancer cases, including carcinoma in situ, arising in the ESECC cohort during the follow-up period from the date of enrollment to December 31, 2016.

2.3. Active follow-up

Active follow-up in this study was conducted in three steps as follows. First, all village doctors of the target villages who are responsible for the primary health care of rural residents were trained by our research team and required to interview all the participants from their villages face-to-face, or at least through direct phone contact if the participants were not at home during the follow-up period. From December 1, 2016, to December 31, 2016, they completed a ‘‘follow-up report form’’ which included personal information for each cohort participant, and for cases of newly diagnosed cancer and death, detailed information regarding the cancer diagnosis and/or death was included. Finally, from January 1, 2017, to April 30, 2017, all 668 community leaders of the target villages were interviewed to confirm the completeness and accuracy of the information reported by the village doctors based on their knowledge of the participants' health history.

2.4. Passive follow-up

Linkage to NCMS claims data was adopted as a passive follow-up method for identification of incident cancer cases. To reduce the impact of prevalent cancers and reimbursement delay on identification of incident cancer cases, NCMS claims data from July 1, 2010, to June 30, 2017, were exported using an annual timeframe based on the date of reimbursement through the Management System of Hua County. Cancer diagnoses were indexed and extracted according to the International Classification of Diseases (10th Version, ICD-10) codes C00 to C97 and D00 to D09 [22].

Claims data for ESECC participants were extracted from the NCMS claims data set using identifiers contained in both data sets, including name and unique personal ID number. Only diagnoses made by secondary (county-level) or tertiary (city-level and above) health care facilities were regarded as valid cancer diagnoses and included in the analysis. For each patient identified, information for the first valid cancer diagnosis in the NCMS data set was used.

2.5. Verification of true incident cancer diagnoses

Cancer diagnoses ascertained using the two follow-up approaches were verified in this study. All cancer cases reported were classified into three groups, including ‘‘both-reported group’’ (cases reported by both active and passive follow-up, not necessarily matched in terms of tumor site and date of diagnosis), ‘‘NCMS-reported group’’ (cases reported in the NCMS data set but not confirmed by active follow-up), and ‘‘AF-reported group’’ (cases reported in active follow-up but not reported by the NCMS data set). Because all NCMS claims data were reported directly by the health facilities where the patients were treated and then rechecked by professionals from NCMS management office, cases in the ‘‘both-reported group’’ were considered to be true cancer cases and were included in the ‘‘verified cancer case data set’’ directly. For these cases, information regarding the tumor site and date of diagnosis in the NCMS data set was adopted. Cases in the AF-reported and NCMS-reported groups were further verified by one-on-one interviews with the patients themselves or their first-degree relatives, or by reviewing a hardcopy of the patients' medical records. Verified true cancer diagnoses were also included in the ‘‘verified cancer case data set.’’

2.6. Interview

To identify potential causes of disagreement between these two follow-up approaches, 32 target village doctors, eight township-level officials, and two staff members of the Hua County NCMS Management Office were interviewed.

2.7. Statistical analysis

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and the Cohen's Kappa coefficient with 95% confidence intervals for active and passive follow-up were calculated using the ‘‘verified cancer case data set’’ as a reference. Stratified analysis of sensitivity was then conducted with selected characteristics of the participants including education, village population size, years of follow-up after enrollment, vital status, and tumor site. ‘‘True cancer diagnosis’’ was defined as a diagnosis which agreed with the ‘‘verified cancer case data set’’ regarding the original site of tumor, where the date of diagnosis was no more than 6 months earlier/later as compared with the date in the ‘‘Verified cancer case data set.’’ People assigned the status ‘‘lost to follow-up’’ in the active follow-up were defined as negative for cancer diagnosis.

All statistical analysis was performed using Stata/SE 14.0 for Windows (STATA Corporation, College Station, TX, USA). All tests were two sided at a significance level of 0.05.

2.8. Ethics statement

Research protocols of the present study were approved by the Institutional Review Board of the Peking University School of Oncology, Beijing, China. All participants provided written informed consent.

3. Results

3.1. Characteristics of study subjects

The characteristics of the participants that may have influenced identification of cancer cases are shown in Table 1. All participants from the 668 target villages in 22 towns of Hua County were between the ages of 45 and 69 at enrollment. Half of these participants (51.3%) were female, and 55.7% had an education level of primary school or below. In addition, most of the participants (71.0%) resided in small- or medium-sized villages with a population size between 500 and 1,499. The median interval of follow-up after enrollment was 1.98 years. A total of 403 (1.2%) participants in this study were lost to active follow-up by December 31, 2016.

Table 1. Selected demographic characteristics for 33,948 participants in the ESECC trial in rural Hua County, China, 2012–2016

Variables	Total (*N* = 33,948) *n* (%)
Age at enrollment, y	
45–59	20,656 (60.8)
60–69	13,292 (39.2)
Gender	
Male	16,543 (48.7)
Female	17,405 (51.3)
Education level[a]	
Primary school or below	18,020 (55.7)
Middle school or above	14,346 (44.3)
Number of towns	22 (100.0)
Number of villages	668 (100.0)
Village population size	
500–1,499	474 (71.0)
1,500–3,000	194 (29.0)
Years of follow-up after enrollment[b]	
Median (quartile)	1.98 (0.92,3.51)
Loss to active follow-up[b]	
Yes	403 (1.2)
No	33,545 (98.8)

Abbreviation: ESECC, Endoscopic Screening for Esophageal Cancer in China.

[a] Education level was based on the 32,336 participants who completed the questionnaire at enrollment.

[b] Years of follow-up after enrollment and loss to active follow-up were calculated from date of enrollment to 31 December 2016.

3.2. Results of active and passive follow-up

In active follow-up, 240 cancer cases were reported by village doctors, and of these, 208 were verified as valid cases and included in the analysis. In addition, another 42 valid cases were reported by community leaders (Fig. 1A). In passive follow-up, 279 subjects in the ESECC cohort were matched in the NCMS data set and 1,165 records with cancer diagnoses were extracted. This included 25 (9.0%) subjects who had more than one cancer diagnosis (Fig. 1B).

3.3. Establishing the "verified cancer case data set"

The flowchart for establishing the "verified cancer case data set" is shown in Figure 2. Among cases reported by active and passive follow-up, 43 were AF-reported, 207 were both-reported, and 72 were NCMS-reported. After the verification process, 12 of the 43 AF-reported and 54 of the 72 NCMS-reported cases were confirmed as true-positive cancer diagnoses. Finally, 273 true cancer cases (12 AF-reported, 207 both-reported, and 54 NCMS-reported) were included in the "verified cancer case data set."

3.4. Sensitivity, specificity, PPV, and NPV of active and passive follow-up

As shown in Table 2, the sensitivity of active follow-up was 54.9% (95% CI: 45.6%–57.4%), which was significantly lower than that of passive follow-up (95.6%, 95% CI: 92.4%–97.7%, $P < 0.001$). The specificity, PPV and NPV of active follow-up were 99.9% (95% CI: 99.8%–99.9%), 82.9% (95% CI: 76.6%–88.1%), and 99.6%

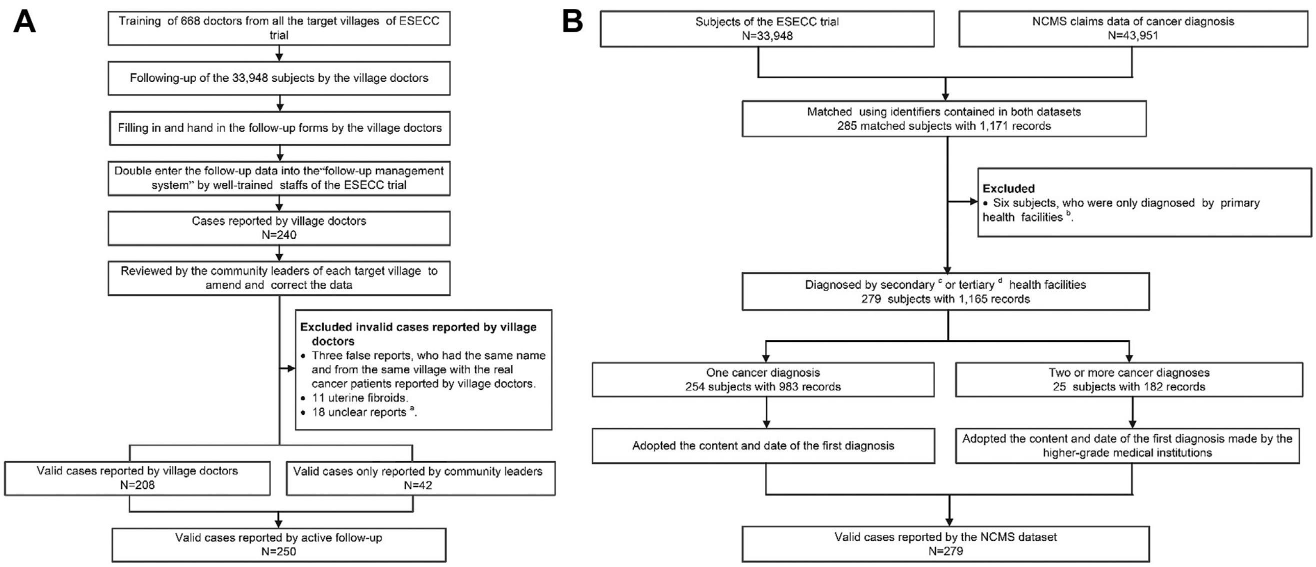
A
Training of 668 doctors from all the target villages of ESECC trial
Following-up of the 33,948 subjects by the village doctors
Filling in and hand in the follow-up forms by the village doctors
Double enter the follow-up data into the"follow-up management system" by well-trained staffs of the ESECC trial
Cases reported by village doctors N=240
Reviewed by the community leaders of each target village to amend and correct the data
Excluded invalid cases reported by village doctors
• Three false reports, who had the same name and from the same village with the real cancer patients reported by village doctors.
• 11 uterine fibroids.
• 18 unclear reports a.
Valid cases reported by village doctors N=208
Valid cases only reported by community leaders N=42
Valid cases reported by active follow-up N=250
B
Subjects of the ESECC trial N=33,948
NCMS claims data of cancer diagnosis N=43,951
Matched using identifiers contained in both datasets 285 matched subjects with 1,171 records
Excluded
• Six subjects, who were only diagnosed by primary health facilities b.
Diagnosed by secondary c or tertiary d health facilities 279 subjects with 1,165 records
One cancer diagnosis 254 subjects with 983 records
Two or more cancer diagnoses 25 subjects with 182 records
Adopted the content and date of the first diagnosis
Adopted the content and date of the first diagnosis made by the higher-grade medical institutions
Valid cases reported by the NCMS dataset N=279

Fig. 1. Flowchart and results of active (A) and passive (B) follow-up in the ESECC trial in rural Hua County, China, 2012–2016. [a]It was unclear whether these cases were reportable or not according to the information provided by village doctors; for example, thyroid lesion, esophageal lesion, and fibroelastoma, etc. [b]Township-level health facilities. [c]County-level health facilities. [d]City-level and above health facilities. *Abbreviation:* ESECC, Endoscopic Screening for Esophageal Cancer in China.

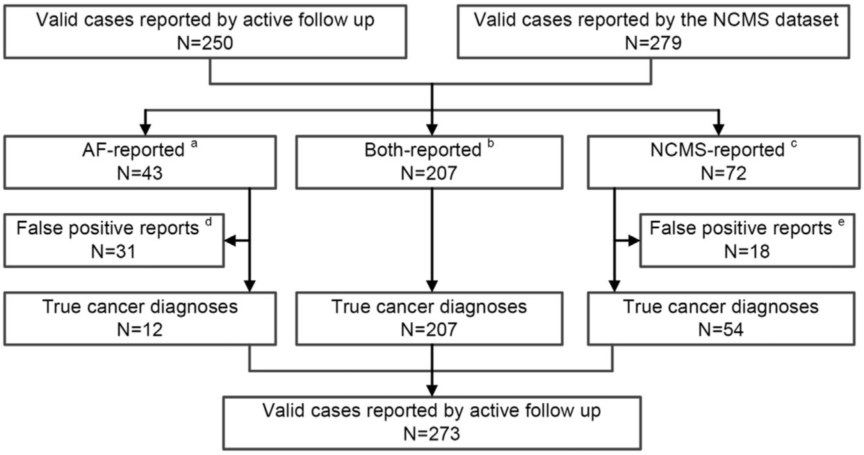
Valid cases reported by active follow up
N=250
Valid cases reported by the NCMS dataset
N=279
AF-reported [a]
N=43
Both-reported [b]
N=207
NCMS-reported [c]
N=72
False positive reports [d]
N=31
False positive reports [e]
N=18
True cancer diagnoses
N=12
True cancer diagnoses
N=207
True cancer diagnoses
N=54
Valid cases reported by active follow up
N=273

Fig. 2. Flowchart for establishment of the ''verified cancer case data set'' in the ESECC trial in rural Hua County, China, 2012–2016. [a]Cases reported by active follow-up but not by the NCMS data set. [b]Cases reported by both active follow-up and the NCMS data set but, which were not necessarily matched as to tumor site and date of diagnosis. [c]Cases reported by the NCMS data set but not by active follow-up. [d]Included 13 begin lesions and 18 nontumor diagnoses. [e]Included four begin lesions and 14 nontumor diagnoses. *Abbreviations:* ESECC, Endoscopic Screening for Esophageal Cancer in China; NCMS, New Rural Cooperative Medical Scheme.

(95% CI: 99.6%–99.7%), respectively. The specificity, PPV, and NPV of passive follow-up were 99.9% (95% CI: 99.9%–100.0%), 93.5% (95% CI: 90.0%–96.1%), and 100.0% (95% CI: 99.9%–100.0%), respectively. The Cohen's Kappa coefficient for active follow-up was 0.66 (95% CI: 0.61–0.71), and for passive follow-up was 0.95 (95% CI: 0.93–0.96).

3.5. Discordance analysis

A total of 154 cases reported by active follow-up showed disagreement with the ''verified cancer case data set'' (Table 3). Among these cases, 54 (35.1%) were unreported, 69 (44.8%) were reported but showed disagreement on the cancer site or the date of diagnosis (Appendix Table A1), and the remaining 31 (20.1%) were false-positive reports. By contrast, 30 cases reported by passive follow-up disagreed with the ''verified cancer case data set,'' including 12 which were unreported by the NCMS data set, and 18 false-positive reports.

3.6. Stratified analysis of sensitivity

Table 4 shows analysis of the sensitivity of the two follow-up strategies stratified by selected characteristics. Sensitivity of active follow-up was higher in males (61.4%) than in females (48.1%, $P = 0.027$) and was significantly higher in common cancers (top five in males and females in the ESECC population, including esophageal, gastric, lung, liver, breast, and colorectal cancer) than in other cancers such as uterus, pancreas, and prostate (67.2% vs. 32.3%, $P < 0.001$). Regarding passive follow-up, sensitivity was high and stable across all the variables.

Table 2. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Cohen's Kappa coefficient of the active and passive follow-up in the ESECC trial in rural Hua County, China, 2012–2016

	Cancer diagnoses							
	True positive	True negative	Total	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Cohen's Kappa coefficient
Data source	**(*N* = 273)**	**(*N* = 33,675)**	**(*N* = 33,948)**	**(%)**	**(%)**	**(%)**	**(%)**	**(95% CI)**
Active follow-up								
Positive	150[a]	31	181	54.9[b] (48.8–60.9)	99.9 (99.8–99.9)	82.9 (76.6–88.1)	99.6 (99.6–99.7)	0.66 (0.61–0.71)
Negative	123	33,644	33,767					
Passive follow-up								
Positive	261	18	279	95.6[b] (92.4–97.7)	99.9 (99.9–100.0)	93.5 (90.0–96.1)	100 (99.9–100.0)	0.95 (0.93–0.96)
Negative	12	33,657	33,669					

Abbreviation: ESECC, Endoscopic Screening for Esophageal Cancer in China.

[a] Correct on tumor site and agreement within ±6 months on date of diagnosis with the ''verified cancer case data set.''

[b] Sensitivity of active follow-up was significantly lower than that of passive follow-up ($P < 0.001$).

Table 3. Discordance analysis of the results of active and passive follow-up with the "verified cancer case data set," in the ESECC trial in rural Hua County, China, 2012–2016

Variables	*N* (%)
Active follow-up[a]	154 (100.0)
Unreported cases	54[b] (35.1)
True-positive cancer cases but disagreement on cancer site or beyond 6 months on date of diagnosis	69 (44.8)
False reports	31[c] (20.1)
Passive follow-up	30 (100.0)
Unreported cases	12[d] (40.0)
False reports	18[e] (60.0)

Abbreviation: ESECC, Endoscopic Screening for Esophageal Cancer in China.

[a] Correct on tumor site and agreement within ±6 months on date of diagnosis with the "verified cancer case data set."

[b] Included 10 true cancer diagnoses were reported as unclear reports by active follow-up and excluded in the quality control process, 41 were never reported by active follow-up, and three were lost to active follow-up.

[c] Included 13 benign lesions and 18 nontumor diagnoses.

[d] Included seven outpatients and five true cancer diagnoses who were diagnosed only by primary health facilities and excluded in the quality control process.

[e] Included four benign lesions and 14 nontumor diagnoses.

4. Discussion

PBCR represents the gold standard for the provision of information on cancer incidence in a defined population. Despite the tremendous advances in PBCRs over the past decade in low- and middle-income countries (LIMCs), registry coverage with high-quality data remained low in LIMCs [5]. The present study was the first study to demonstrate the power of health reimbursement data for identifying incident cancers in less developed areas without PBCRs, where active door-to-door interview was the only choice available, and interviewing had been adopted years earlier in many large-scale cohort studies. The results of the present study demonstrated a much more efficient strategy for tracking the occurrence of cancer in prospective cohorts and establishing PBCRs in less developed areas.

In our study, we found that the overall sensitivity of linking with NCMS claims data was fairly high (95.6%). This finding is consistent with previous studies, which had shown in other countries that claims data from health insurance systems could be used to identify incident cancer cases with ideal sensitivity and specificity as compared to cancer registries data [7–19]. For example, Jonathan et al. found a high sensitivity (95%) and specificity (97%) of Medicare claims data in identifying patients with oral and pharyngeal cancer [17]. Izumi et al. found that breast cancer cases could be accurately identified from the Japanese claims database using an optimal procedure, which had high sensitivity (90.4%) and specificity (99.8%) [19].

In rural China, the NCMS has been the most important, and under some circumstances, the only medical insurance that local people have. This insurance plays a leading role in improving the rural population's access to health services and alleviating poverty due to catastrophic health expenses [23]. Prompted by economic benefit, almost all cancer patients apply for reimbursement from NCMS in rural China, which ensures an extremely low risk of missing cases. Another advantage of using claims data to ascertain cancer cases was that it would include diagnostic and treatment information regardless of where diagnosis or treatment was rendered. This enabled us to capture incident cancer patients who might have otherwise been missed by traditional active follow-up due to population migration caused by the economic development in today's China, which has been an increasing challenge to population-based prospective studies [24]. Moreover, NCMS claims data contain detailed personal information, date of admission and discharge, cancer diagnosis, International Classification of Diseases (10th Version, ICD-10) code, and medical expenses, all recorded in a standard and routine manner, which provides valuable detailed records on patients' diagnoses and treatments [23]. Thus, NCMS claims data were used to identify incident cancer cases with high sensitivity in this study.

By contrast, our results indicated that the overall sensitivity of active follow-up was relatively low (54.9%), especially when considering the accuracy of the reported tumor site and date of diagnosis. In previous studies taking cancer registry data as a reference, the sensitivity of self-reported diagnosis varied from 53.0% in Japan, to 71.1% in Australia, and to 86.6% in the United States [25–27]. This may be due to the fact that cancer disclosure is heavily influenced by cultural context [25]. In rural China, patients were sometimes not informed of a cancer diagnosis, and due to conservative cultural attitudes, family members were reluctant to let others know about occurrence of cancer [28]. Thus, it was difficult to obtain complete and accurate information regarding cancer diagnosis directly from patients themselves or from family members in our study population. Although the ESECC trial had committed substantial resources to the active follow-up process [21], data provided by village doctors still did not reach the standards of the NCMS data.

The higher sensitivity of active follow-up observed in men in this study might be due in part to the fact that rural China is still a male-dominated society [29], and males are better known to village doctors than females according to our interviews. Male cancer patients were therefore more likely to be identified and recorded. Consistent with previous studies [25,30], we also found sensitivity of active follow-up varied considerably with cancer site. The principal reason was that common cancers with high prevalence such as esophageal cancer, stomach cancer, or lung cancer were also better recognized by people and more likely to be reported than rare cancers.

Table 4. Sensitivity of active and passive follow-up by selected characteristics in the ESECC trial in rural Hua County, China, 2012–2016

	Verified cancer cases	Active follow-up			Passive follow-up		
		True positive	Sensitivity[a]		True positive	Sensitivity	
Variables	*n* (%)	*n* (%)	(%)	*P* value[b]	*n* (%)	(%)	*P* value[b]
Total	273 (100.0)	150 (100.0)	54.9		261 (100.0)	95.6	
Age group (years)							
45–59	124 (45.4)	67 (44.7)	54.0	0.782	120 (46.0)	96.8	0.390
60–69	149 (54.6)	83 (55.3)	55.7		141 (54.0)	94.6	
Gender							
Male	140 (51.3)	86 (57.3)	61.4	0.027	133 (51.0)	95.0	0.617
Female	133 (48.7)	64 (42.7)	48.1		128 (49.0)	96.2	
Education level[c]							
Primary school or below	158 (62.0)	87 (62.6)	55.1	0.821	149 (61.1)	94.3	0.165
Middle school or above	97 (38.0)	52 (37.4)	53.6		95 (38.9)	97.9	
Village population size							
500–1,499	130 (47.6)	65 (43.3)	50.0	0.117	125 (47.9)	96.2	0.673
1,500–3,000	143 (52.4)	85 (56.7)	59.4		136 (52.1)	95.1	
Years of follow-up after enrollment [d]							
0-1	54 (19.8)	34 (22.7)	63.0	0.186	53 (20.3)	98.1	0.301
2-4	219 (80.2)	116 (77.3)	53.0		208 (79.7)	95.0	
Vital status[e]							
Living	169 (61.9)	96 (64.0)	56.8	0.431	164 (62.8)	97.0	0.140
Dead	104 (38.1)	54 (36.0)	51.9		97 (37.2)	93.3	
Tumor site							
Common cancers[f]	177 (64.8)	119 (79.3)	67.2	<0.001	167 (64.0)	94.4	0.170
Other cancers	96 (35.2)	31 (20.7)	32.3		94 (36.0)	97.9	

Abbreviation: ESECC, Endoscopic Screening for Esophageal Cancer in China.

[a] Correct on tumor site and agreement within ±6 months on date of diagnosis with the ‘‘verified cancer case data set.’’

[b] *P* values were derived from the chi-square test.

[c] The education level was based on the 32,336 participants who completed the questionnaire at enrollment.

[d] The years of follow-up after enrollment were calculated from date of enrollment to December 31, 2016.

[e] The vital status was evaluated by the end of 31st December 2016.

[f] Top five cancers for male and female in the ESECC cohort, including cancer of the esophagus, stomach (included cardia), lung, liver, breast, and colorectal cancer.

Because population-level follow-up is labor intensive and time consuming, cost must be taken into consideration when designing the working protocol. Active follow-up in the ESECC trial involved 668 village doctors, 22 township leaders, and nine full-time investigators in our research center in Hua County to ensure a high rate of response and data quality. About 5 months were spent to complete the first round of active follow-up process. By contrast, only five staff members and about 21 days were needed for passive follow-up. As such, passive linkage with health insurance system claims data was superior in terms of resource consumption as compared to active follow-up.

Use of NCMS claims data to identify cancer cases has a limitation, which should be noted. Theoretically, the NCMS system does not record data from outpatients and uninsured cancer cases. However, for a fatal chronic disease like cancer, which often requires multidisciplinary therapy in high-level hospitals and is only rarely treated in primary facilities, the proportion of outpatients was extremely low. For example, only seven (2.6%) cancer cases in this study were diagnosed and treated in outpatient clinics. In addition, the Chinese central government has invested an enormous amount of resources to insure all rural residents are covered by the NCMS, and the coverage of NCMS has reached almost 100% in rural China by 2013 [23].

5. Conclusion

In summary, passive linkage with claims data from the NCMS system was an efficient approach for identifying incident cancer cases in a rural population in China. Compared to traditional door-to-door active follow-up, linkage with a government-run health insurance system should be extensively used in regions and populations without specific disease registries like PBCRs to track the occurrence of major chronic diseases, which are generally diagnosed and treated in high-level hospitals.

CRediT authorship contribution statement

Chao Shi: Investigation, Writing - original draft, Formal analysis. **Mengfei Liu:** Investigation, Formal analysis. **Zhen Liu:** Investigation, Formal analysis. **Chuanhai Guo:** Investigation. **Fenglei Li:** Investigation. **Ruiping Xu:** Investigation. **Fangfang Liu:** Investigation. **Ying Liu:** Investigation. **Jingjing Li:** Investigation. **Hong Cai:** Conceptualization, Writing - original draft. **Zhonghu He:** Conceptualization, Methodology, Investigation, Writing - original draft, Formal analysis, Supervision. **Yang Ke:** Conceptualization, Methodology, Writing - original draft, Supervision.

Acknowledgments

The authors thank all the following team members and collaborators for their contributions to the fieldwork done for this study including 668 village doctors in the target villages of the ESECC trial, and 22 township leaders. The authors also thank the nine employed investigators in their research center in Hua County including Lijuan An, Jiaoyan Feng, JunyingLv, Jiaqi Duan, Hongxing Zhu, Ruifang Li, Ying Chen, Xiangyun Fang, and Sumin Qi, and Liang Lei from the NCMS Management Office of Hua County. The authors would also like to thank the Government of Hua County, Health and Family Planning Commission of Hua County, Henan Province, and all the participants of the ESECC trial. The authors would like to thank Dr. Michael A. McNutt for editing this manuscript.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.06.009.

References

[1] Kato I, Toniolo P, Koenig KL, Kahn A, Schymura M, Zeleniuch-Jacquotte A. Comparison of active and cancer registry-based follow-up for breast cancer in a prospective cohort study. Am J Epidemiol 1999;149:372−8.

[2] Valsecchi MG, Steliarova-Foucher E. Cancer registration in developing countries: luxury or necessity? Lancet Oncol 2008;9:159−67.

[3] Rashbass J, Peake M. The evolution of cancer registration. Eur J Cancer Care 2014;23:757−9.

[4] Wei KR, Chen WQ, Zhang SW, Liang ZH, Zheng RS, Ou ZX. Cancer registration in the peoples Republic of China. Asian Pac J Cancer Prev 2012;13:4209−14.

[5] Bray F, Znaor A, Cueva P, Korir A, Swaminathan R, Ullrich A, et al. Planning and developing population -based cancer registration in low- and middle incomes setting. Lyon, France: International Agency For Research On Cancer; 2014.

[6] Chatterjee S, Chattopadhyay A, Senapati SN, Samanta DR, Elliott L, Loomis D, et al. Cancer Registration in India - current scenario and future perspectives. Asian Pac J Cancer Prev 2016;17:3687−96.

[7] McClish DK, Penberthy L, Whittemore M, Newschaffer C, Woolard D, Desch CE, et al. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. Am J Epidemiol 1997;145:227−33.

[8] Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. Med Care 1999;37:436−44.

[9] Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. J Clin Epidemiol 2000;53:605−14.

[10] Wang PS, Walker AM, Tsuang MT, Orav EJ, Levin R, Avorn J. Finding incident breast cancer cases through US claims data and a state cancer registry. Cancer Causes Control 2001;12:257−65.

[11] Koroukian SM, Cooper GS, Rimm AA. Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population. Health Serv Res 2003;38:947−60.

[12] Barzilai DA, Koroukian SM, Neuhauser D, Cooper KD, Rimm AA, Cooper GS. The sensitivity of Medicare data for identifying incident cases of invasive melanoma (United States). Cancer Causes Control 2004;15:179−84.

[13] McClish D, Penberthy L. Using Medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. Med Care 2004;42:1111−6.

[14] Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. Health Serv Res 2004;39:1733−49.

[15] Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in Medicare claims data. Health Serv Res 2007;42: 2056−69.

[16] Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. J Clin Epidemiol 2008;61: 373−9.

[17] Mahnken JD, Keighley JD, Girod DA, Chen X, Mayo MS. Identifying incident oral and pharyngeal cancer cases using Medicare claims. BMC Oral Health 2013;13:1.

[18] Ajrouche A, Estellat C, De Rycke Y, Tubach F. Evaluation of algorithms to identify incident cancer cases by using French health administrative databases. Pharmacoepidemiol Drug Saf 2017;26: 935−44.

[19] Sato I, Yagata H, Ohashi Y. The accuracy of Japanese claims data in identifying breast cancer cases. Biol Pharm Bull 2015;38:53−7.

[20] Notice of the Ministry of Science and Technology Issuing the Guidelines for the 2016 Project declaration on precision medical research of National Science and Technology Major Project. People's Republic of China: Ministry of Science and Technology; 2016.

[21] He Z, Liu Z, Liu M, Guo C, Xu R, Li F, et al. Efficacy of endoscopic screening for esophageal cancer in China (ESECC): design and preliminary results of a population-based randomised controlled trial. Gut 2019;68:198−206.

[22] Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). 2016. Available at https://icd.who.int/browse10/2016/en. Accessed March 18, 2019.

[23] Li X, Cai H, Wang C, Guo C, He Z, Ke Y. Economic burden of gastrointestinal cancer under the protection of the New Rural Cooperative Medical Scheme in a region of rural China with high incidence of oesophageal cancer: cross-sectional survey. Trop Med Int Health 2016;21:907−16.

[24] Mues KE, Liede A, Liu J, Wetmore JB, Zaha R, Bradbury BD, et al. Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. Clin Epidemiol 2017;9:267−77.

[25] Inoue M, Sawada N, Shimazu T, Yamaji T, Iwasaki M, Sasazuki S, et al. Validity of self-reported cancer among a Japanese population: recent results from a population-based prospective study in Japan (JPHC Study). Cancer Epidemiol 2011;35:250−3.

[26] Loh V, Harding J, Koshkina V, Barr E, Shaw J, Magliano D. The validity of self-reported cancer in an Australian population study. Aust N Z J Public Health 2014;38:35−8.

[27] Pinsky PF, Yu K, Black A, Huang WY, Prorok PC. Active follow-up versus passive linkage with cancer registries for case ascertainment in a cohort. Cancer Epidemiol 2016;45:26−31.

[28] Sun W, Wang Z, Fang S, Li M. Factors influencing the attitudes of Chinese cancer patients and their families toward the disclosure of a cancer diagnosis. J Cancer Educ 2015;30: 20−5.

[29] Hou F, Cerulli C, Wittink MN, Caine ED, Qiu P. Depression, social support and associated factors among women living in rural China: a cross-sectional study. BMC Womens Health 2015;15:28.

[30] Navarro C, Chirlaque MD, Tormo MJ, Perez-Flores D, Rodriguez-Barranco M, Sanchez-Villegas A, et al. Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. J Epidemiol Community Health 2006;60:593−9.