ELSEVIER

Seminars in
RADIATION
ONCOLOGY

Use of Claims Data for Cost and Cost-Effectiveness Research

Ya-Chen Tina Shih, PhD,* and Lei Liu, PhD†

Administrative claims data are big data generated from healthcare encounters. Claims data contain information on insurance payment as well as clinical diagnoses and procedure codes to ascertain medical conditions and treatments, making them valuable sources for economic evaluation research. This paper offers an introductory overview of the use of claims data for oncology-related cost-of-illness, cost comparison, and cost-effectiveness analyses. We reviewed analytical methods commonly employed in these analyses, such as the phase of care approach and net costing method for cost-of-illness studies, propensity score matching methods for cost comparison studies, and net benefit regression models for cost-effectiveness studies. We used published studies to explain each method and to discuss methodological challenges of conducting economic studies using claims data.
Semin Radiat Oncol 29:348–353

Introduction

The complexity of modern medicine produces a massive amount of data at each medical care encounter, from body measurement and laboratory test readings to health outcomes and payment associated with medical services received during the encounter. One type of big data generated from healthcare encounters that are valuable for economic evaluation research are administrative claims data. These data gather information on services received by patients enrolled in a health insurance plan, diagnoses related to these services, and costs associated with each service item.

In the United States, claims data can be obtained from public and private insurance or healthcare systems. Commonly used claims data from public insurance are Medicare and Medicaid claims. A further enhancement of Medicare claims data is the linked Surveillance, Epidemiology, and End Results (SEER)-Medicare claims data.[1] The SEER-Medicare data provide both clinical and economic information for elderly cancer patients and have been the primary resource for oncology health services research since its inception. Claims data from private insurance are often proprietary and can be licensed from commercial vendors, such as, MarketScan data distributed by IBM Corporation, LifeLink data by the IMS Health, and Optum data by Optum Life Sciences. Examples of claims data from healthcare systems are data from Kaiser or Blue Cross Blue Shield. Outside the United States, national claims data are available for several countries with national health insurance programs, such as Taiwan, Sweden, and Korea.[2] A detailed overview of large databases, along with their strengths and limitations, can be found in Jagsi et al (2014).[3]

Economic evaluation research using claims data typically encompasses 2 types of studies: cost and cost-effectiveness analysis (CEA). Payment information recorded in claims data form the basis for cost analyses whereas the combination of payment and outcomes information extracted or derived from claims offers useful data for cost-effectiveness analyses. In this article, we discussed the use of administrative claims data in oncology-related cost and cost-effectiveness analyses. We drew examples from published studies to explain analytical methods used in these analyses and discuss methodological challenges commonly encountered in this line of research.

Use of Claims Data for Cost Analyses

Claims data are used in 2 types of cost analyses: cost-of-illness (COI) and cost comparison studies. COI studies estimate the economic burden of a specific disease or medical

*Section of Cancer Economics and Policy, Department of Health Services Research, The University of Texas M. D. Anderson Cancer Center, Houston, TX
†Division of Biostatistics, Washington University School of Medicine in St. Louis, MO
We acknowledge funding from the National Cancer Institute (Shih, R01CA207216, R01CA225646, and CCSG P30 CA016672)
Conflict of Interest: None.
Address reprint requests to Ya-Chen Tina Shih, PhD, Section of Cancer Economics and Policy, Department of Health Services Research, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. Univ 1444, Houston TX 77030. E-mail: yashih@mdanderson.org

https://doi.org/10.1016/j.semradonc.2019.05.009
1053-4296/

condition whereas cost comparison studies assess treatment costs and associated downstream costs between therapeutic alternatives.

COI Studies

COI studies are important for budget planning purposes as they inform policy makers of the financial burden of specific illnesses. Among a wide array of illnesses, costs of cancer care are of immerse interest to policy makers because of the rapidly growing cost of cancer treatments and the anticipated rising incidence of cancer due to aging demographics. Researchers at the National Cancer Institute have routinely used the linked SEER-Medicare data to estimate costs of cancer care.[4-6] Estimates from this research contribute to the Financial Burden of Cancer Care section of the Cancer Trends Progress Reports released by the National Cancer Institute annually.[7] For example, the latest estimates reported that national medical costs of cancer were $124.6 billion in 2010 and projected to be $157.8 billion by 2020.[8]

Two analytical approaches commonly used to estimate the costs of cancer care are the prevalence and incidence approach. The prevalence approach reports costs of cancer for a specific time period, whereas the incidence approach follows a newly diagnosed cohort to keep track of costs throughout the cancer care continuum.[9,10] Each costing approach provides different but equally important information to policy makers. Estimates of prevalence costs are helpful in that knowledge of cancer care costs in the current and previous calendar years can be used to project future healthcare expenditures, whereas incidence costs inform policy makers regarding the effect of cancer prevention, treatment, or intervention on the cost trajectory of an incident cohort of cancer patients.[10]

Studies reporting cost of cancer care often stratified the cost estimates by phase of care, with the initial care phase covering the first 12 months following cancer diagnosis, the terminal care phase occupying the last 12 months of life, and the continuing care phase capturing all the months in between. An analytical challenge is to determine the extent to which the costs data captured in the claims of cancer patients were attributable to cancer care. This is accomplished by employing the incremental costing method (also known as net costing method) in which a matched control cohort of noncancer patients was constructed for a cohort of cancer patients to form the basis of what costs would be had these patients not had cancer. Matching factors commonly found in the literature included age (often at 5-year intervals), sex (for cancers that are not gender specific), race, and SEER area strata.[4,5] Cancer-related costs (or "net costs") are then calculated as the difference between the mean cost of the cohort of cancer patients and that of the matched noncancer control. Table 1 describes the net costs of breast cancer patients from a recent analysis of SEER-Medicare data.[11] It shows that the net costs were higher for initial and terminal care phases and tended to be higher for cancer diagnosed at more advanced stages. It is worth noticing that the reporting of terminal care phase costs made a distinction between cancer patients who had cancer as the cause of death vs those who did not. The logic behind such a distinction is that the

Table 1 Net Costs of Breast Cancer by Phase of Care, Cancer Stage, and First-Year Treatment Pattern (US 2017 dollars)

	Stage 0	Stage I	Stage II	Stage III	Stage IV
Initial care phase					
No chemo, no RT, no surgery	$0	$692	$8029	$13,645	$23,009
No chemo, no RT, lumpectomy	$3663	$8029	$11,244	$16,170	$27,106
No chemo, no RT, mastectomy	$10,819	$12,246	$15,737	$22,304	$31,988
No chemo, RT, no surgery	$5546	$5015	$10,913	$24,106	$41,584
No chemo, RT, lumpectomy	$15,633	$19,110	$21,803	$27,460	$36,420
No chemo, RT, mastectomy	$19,783	$21,073	$26,923	$31,061	$39,258
Chemo, no RT, no surgery	$7784	$7579	$28,439	$32,568	$54,879
Chemo, no RT, lumpectomy	$10,994	$28,660	$35,240	$55,159	$71,826
Chemo, no RT, mastectomy	$26,173	$36,826	$41,012	$53,042	$62,846
Chemo, RT, no surgery	$19,856	$32,274	$38,933	$55,059	$77,786
Chemo, RT, lumpectomy	$18,473	$42,913	$49,279	$58,420	$64,904
Chemo, RT, mastectomy	$26,088	$51,413	$55,631	$62,267	$73,118
Continuing care phase					
Year 1	$0	$1270	$3413	$8361	$23,278
Year 2	$0	$774	$2268	$5696	$20,811
Year 3	$0	$747	$2147	$5372	$20,042
Year 4	$0	$943	$2438	$4276	$17,674
Year 5	$0	$669	$1790	$3004	$13,054
Year 6+	$0	$639	$1107	$2968	$13,446
Terminal care phase					
Cause of death: breast cancer	$41,822	$47,824	$51,228	$57,186	$70,603
Cause of death: other	$7321	$3263	$5221	$11,715	$31,170

Abbreviation: RT, radiation therapy.
Note: Initial care was defined as care incurred within the first 12 months of diagnosis, terminal care reflected the last 12 months of life, and continuing care captured everything that happens between initial and terminal care phases.
Data adapted from Shih et al (2019).

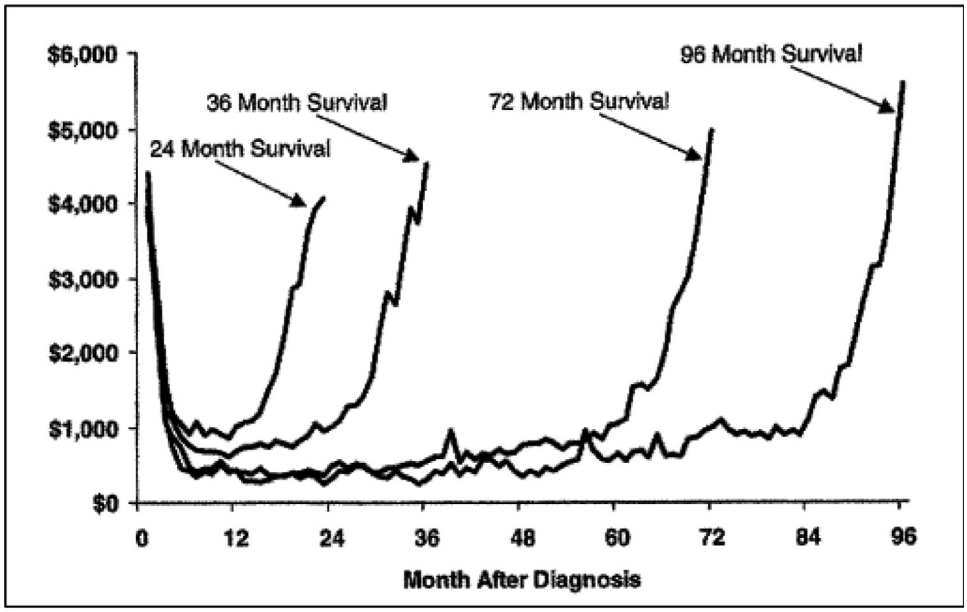

Figure 1 Average monthly costs of cancer, by survival time. Data adapted from Brown et al (2002).

end-of-life care for cancer patients who did not die of cancer likely resembles the care for their counterparts in the matched noncancer cohort; thus the net costs in terminal care phase for these patients would be lower than for patients who die from cancer.

A more sophisticated way to present cost information is to show the cost trajectory throughout patients' lifetime. Not surprisingly, research examining the longitudinal cost of cancer patients often reported that the cost trajectory of cancer care exhibited a U-shape curve as treatment tended to be most intense at the beginning as well as toward the end of the cancer care continuum. Figure 1 illustrates the cost trajectory of patients whose breast cancer was diagnosed between 1980 and 1998 based on an earlier analysis of SEER-Medicare data.[6] Analyses fitting cost trajectories with cross-sectional data while neglecting the longitudinal cost data structure could lead to inefficient and sometimes inaccurate inferences as important information, such as the time pattern of critical medical events, can be lost in models using snapshots of cross-sectional data. More advanced and statistically rigorous analyses of cost trajectories should consider the correlation between monthly (quarterly or yearly) costs with the longitudinal data structure[12] and also account for censoring.[13]

Cost Comparison Studies

When it is feasible to differentiate treatments using billing codes such as the Current Procedural Terminology or ICD-9 or ICD-10 procedure codes, claims data can provide useful information to compare costs between treatments. In addition, because claims data collect information on the date medical services rendered, researchers can construct a clinically meaningful time period (eg, 3 months) subsequent to the treatment receipt date to keep track of downstream events, such as treatment-related complications, and the associated costs. This information is important in assessing the cost impact of new medical technologies as new treatments are almost always more expensive than current standard of care and drug or medical device companies often cite better health outcomes (eg, fewer complications, better survival) to justify their higher costs. Therefore, cost comparison studies that compared treatment costs alone provide limited information to decision makers.

The study by Pan et al (2018) offered an example of more comprehensive cost comparison studies.[14] In this study, the authors analyzed 2008-2015 MarketScan Commercial Claims and Encounters database to compare toxicities and cost of 3 forms of radiation therapy, proton radiation, stereotactic body radiotherapy (SBRT), and intensity-modulated radiotherapy (IMRT), for privately insured prostate cancer patients under the age of 65. Radiation therapy was identified from claims data using a combination of Current Procedural Terminology and ICD-9 procedure codes. Using the first date indicative of radiation as the index date, the authors defined the receipt of radiation therapy as having at least 3 fractions of SBRT or 20 fractions of IMRT or proton radiation within 90 days of the index date and modeled cumulative incidence of various types of toxicities and the associated costs at 6, 12, 24, and 36 months after the index date. The comparison between IMRT and proton therapy showed that at 2 years total healthcare cost of proton therapy was substantially higher than that of IMRT ($133,220 vs $79,209; $P < 0.001$), although a portion of the higher treatment cost was offset by lower complication costs ($1737 vs $2730; $P = 0.008$). The comparison between SBRT and IMRT showed similar mean total cost ($80,786 vs $77,539; $P = 0.36$) and complication cost ($3084 vs $2079; $P = 0.25$) at 2 years.

An analytical strategy employed in this paper warranted more discussions. The authors used propensity score-matched case-control study design to balance the observable covariates between treatment groups. This adjustment is important because it created more comparable cohorts of patients to compare the costs and downstream events between treatments. For example, if healthier patients were more likely to receive proton therapy, such favorable selection could lead to a better toxicity profile, which then translates to lower complication costs for patients in the proton therapy group. The use of a propensity scored-matched method helped mitigate biases arisen from treatment selection. Alternatively, one could conduct regression analysis to adjust for covariate effects on costs and report results of cost comparison as the "adjusted" difference.

An example of this approach can be found in Guadagnolo et al (2013).[15] In this article, the authors analyzed claims data in the SEER-Medicare database for over 200,000 patients who died of lung, breast, prostate, colorectal, and pancreas cancers between 2000 and 2007 to compare end-of-life Medicare payment for 4 groups of cancer patients classified by whether the patient received radiation therapy (yes vs no) and/or hospice care (yes vs no) in the last month of life. The study found that costs were highest for the group of patients who received end-of-life radiation therapy but did not receive hospice care, the mean "adjusted" cost for this group was $3453 (95% CI: $3176 − $3730, in 2009 US dollars) higher than that of the group of patients who did not receive radiation therapy nor hospice care in the last month of life. Without covariate adjustment, the "unadjusted" cost difference was $2483 (95% CI: $2092 − $2874). The use of regression-based methods in cost comparison studies often encounters a statistical issue that the medical cost data are highly skewed to the right because a small number of patients tend to consume a rather large proportion of healthcare expenditures. To deal with this issue, the authors applied an econometric technique called the extended estimating equations method.[16]

It should be noted that even after applying the propensity score method or the regression-based approach, there could still be unobserved factors contributing to treatment selections (eg, preference toward newer technologies among well-informed patients). The econometric literature recommends the use of instrument variables (IV) method to address the issue of selection bias. However, successful implementations of this method rely critically on finding the appropriate IVs, which is often challenging empirically. Further, with weak IVs, the performance of the IV method is worse than the standard least squared models, as demonstrated in the study by Hadley et al (2003).[17] The authors analyzed Medicare claims to compare 3-year survival of 3 treatments for women with early stage breast cancer: mastectomy, breast conserving surgery plus radiation therapy, and breast conserving surgery alone. Using both ordinary least squares and IV methods, this study demonstrated that if the IVs were weak, estimates from the IV method would not only be biased but also inconsistent, which could result in misleading conclusion in hypothesis testing. Given the practical and technical difficulties in executing IV methods, it is our opinion that when it is not feasible to apply IV methods in cost comparison studies, researchers should present findings from both unadjusted and adjusted analyses to better understand the impact of observable covariates.

Use of Claims Data for Cost-Effectiveness Analyses

CEA has been referred to as the fourth hurdle for new technologies and interventions, following the traditional 3 hurdles for licensing requirements: safety, efficacy, and quality.[18] CEA helps decision-makers understand the trade-offs between costs and health outcomes associated with new interventions.[19] Conventional CEA reports study findings in terms of the incremental cost-effectiveness ratio (ICER).[19,20] The ICER, calculated as the difference in mean costs between the new and standard treatment divided by the difference in mean effectiveness between the 2, estimates additional resources needed to achieve an increase in units of effectiveness. The ICER is then compared with a threshold value (such as $100,000 per QALY) to determine whether a new treatment is cost effective.

The analytical approach employed in CEA depends on the type of data available. Overall, a statistical approach is applied when patient-level data are available from clinical trials or observational data, whereas a modeling approach is used to synthesize information from a mix of data sources, such as published literature, patient-level data, and expert opinions.[21] Claims data contribute to CEA indirectly or directly. Indirectly, costs estimated from claims data can be incorporated into models designed for CEA. For example, the stage and phase of care-specific net costs of breast cancer estimated from SEER-Medicare data (see Table 1 above) were used to populate the treatment cost portion of a micro-simulation model to assess the cost effectiveness of breast cancer screening guidelines from different professional societies.[11] Directly, researchers can obtain both health outcomes and costs information of comparators from claims data, and use such information to estimate incremental costs and incremental effectiveness for the calculation of the ICER.

An example of claims data-based CEA can be found in Shaya et al (2014).[22] The authors constructed an incident cohort of hepatocellular carcinoma (HCC) patients diagnosed between 2000 and 2007 from SEER-Medicare and compared the cost effectiveness of HCC treatment modalities by stage. Treatment modalities considered in the analysis included: transplant, resection, liver-directed therapy, radiation, chemotherapy, and no treatment. "Effectiveness" was measured as years survived and estimated using Cox proportional hazards model. "Cost" was quantified as cumulative Medicare expenditures and the authors applied the partitioned inverse probability weighted method to account for right-censored data. This study concluded that resection was the most cost-effective treatment modality for early stage or unstaged HCC patients, whereas liver-directed therapy was more cost effective than chemotherapy or radiation for stage

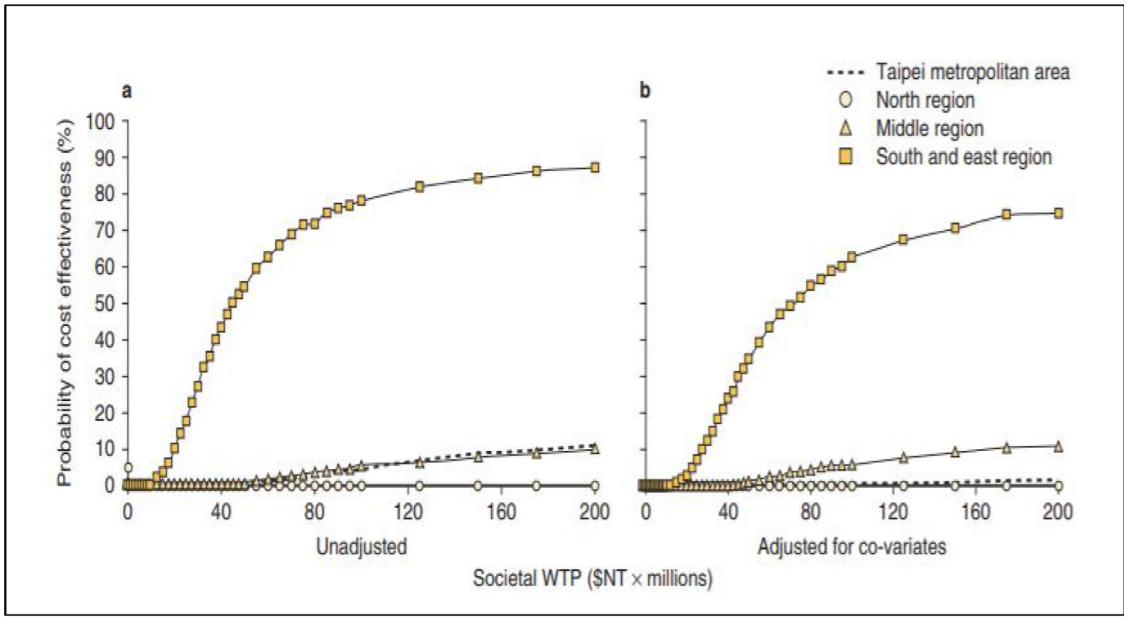

Figure 2 Cost-effectiveness acceptability curve of CEF vs CMF by geographic regions in Taiwan (in $NT 2006).
Note: CEF, cyclophosphamide, epirubicin, fluorouracil; CMF, cyclophosphamide, methotrexate, fluorouracil; NT, new Taiwanese dollars ($NT); WTP, willingness to pay.
Data adapted from Shih et al (2009).

IV patients. A methodological concern was that the ICER reported in this study did not consider the correlation between costs and effectiveness as it was obtained from the incremental costs and incremental effectiveness estimated from regression models separately, not jointly.

More advanced CEA will transform ICER into net benefit, defined as $NB(\lambda) = \lambda \bullet \Delta E - \Delta C$, where λ represents a societal willingness-to-pay, ΔC represents the incremental costs, and ΔE represents the incremental effectiveness, and report findings as the cost-effectiveness acceptability curve.[23,24] The cost-effectiveness acceptability curve informs decision makers of the probability that the new intervention is more cost effective than the standard treatment corresponds to various levels of societal willingness-to-pay. An advantage of the NB transformation is that the NB can be incorporated into a regression framework to allow for covariate adjustments and the examination of interaction effects.[25] Another advantage is that the correlation between costs and effectiveness is automatically considered in this transformation.

An application of the NB regression framework to claims data can be found in Shih et al (2009).[26] The authors applied this method to breast cancer patients identified from claims data in the National Health Insurance Research Database in Taiwan to compare the cost effectiveness of 2 commonly prescribed first-line chemotherapy regimens for breast cancer patients in Taiwan: cyclophosphamide, methotrexate, fluorouracil vs cyclophosphamide, epirubicin, fluorouracil. NB_i for each patient was constructed as $\lambda \bullet E_i - C_i$ and used as the dependent variable in the NB regression. In addition to treatment, other covariates included in the NB regression model were age, geographic region, type of surgery (mastectomy vs lumpectomy), facility type, and comorbidities. Findings from this study are summarized in Figure 2. It indicates (a) a strong interaction effect between treatment cost effectiveness and geographic region, and (b) the cost effectiveness of cyclophosphamide, methotrexate, fluorouracil vs cyclophosphamide, epirubicin, fluorouracil was sensitive to covariate adjustments. More in-depth discussions of the application of regression-based approaches in patient-level data for CEA are provided in Goto et al (2017).[27]

Concluding Remarks

With advances in information technology in automated data collection and management techniques, claims data are expected to become available more quickly. In addition, increasing availability of genetic and biomarker data can further enrich the content of claims data with clinical information gathered in electronic medical records. To ensure the analytical rigor and the accuracy of statistical inferences, researchers who are interested in using claims data for economic evaluation research should become familiar with methods discussed in this paper and understand the strengths and limitations of these methods when interpreting their study findings.

References

1. Warren JL, Klabunde CN, Schrag D, et al: Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 40:IV-3-18, 2002
2. Hsing AW, Ioannidis JP: Nationwide population science: lessons from the Taiwan national health insurance research database. JAMA Intern Med 175:1527-1529, 2015

3. Jagsi R, Bekelman JE, Chen A, et al: Considerations for observational research using large data sets in radiation oncology. Int J Radiat Oncol Biol Phys 90:11-24, 2014
4. Mariotto AB, Yabroff KR, Shao Y, et al: Projections of the cost of cancer care in the United States: 2010-2020. J Natl Cancer Inst 103:117-128, 2011
5. Yabroff KR, Lamont EB, Mariotto A, et al: Cost of care for elderly cancer patients in the United States. J Natl Cancer Inst 100:630-641, 2008
6. Brown ML, Riley GF, Schussler N, et al: Estimating health care costs related to cancer treatment from SEER-Medicare data. Med Care 40: IV104-117, 2002
7. NCI: cancer trends progress reports, 2018. https://progressreport.cancer.gov/. Accessed February 8, 2019
8. Mariotto AB, Yabroff KR, Feuer EJ, et al: Projecting the number of patients with colorectal carcinoma by phases of care in the US: 2000-2020. Cancer Causes Control 17:1215-1226, 2006
9. Lipscomb J, Yabroff KR, Brown ML, et al: Health care costing: data, methods, current applications. Med Care 47:S1-S6, 2009
10. Barlow WE: Overview of methods to estimate the medical costs of cancer. Med Care 47:S33-S36, 2009
11. Shih YC, Dong W, Xu Y, et al: Assessing the cost-effectiveness of updated breast cancer screening guidelines for average-risk women. Value Health 22:185-193, 2019
12. Chen J, Liu L, Zhang D, et al: A flexible model for the mean and variance functions, with application to medical cost data. Stat Med 32:4306-4318, 2013
13. Li L, Wu CH, Ning J, et al: Semiparametric estimation of longitudinal medical cost trajectory. J. Am. Stat Assoc 113:582-592, 2018
14. Pan HY, Jiang J, Hoffman KE, et al: Comparative toxicities and cost of intensity-modulated radiotherapy, proton radiation, and stereotactic body radiotherapy among younger men with prostate cancer. J Clin Oncol 36:1823-1830, 2018
15. Guadagnolo BA, Liao KP, Elting L, et al: Use of radiation therapy in the last 30 days of life among a large population-based cohort of elderly patients in the United States. J Clin Oncol 31:80-87, 2013
16. Basu A, Arondekar BV, Rathouz PJ: Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. Health Econ 15:1091-1107, 2006
17. Hadley J, Polsky D, Mandelblatt JS, et al: An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. Health Econ 12:171-186, 2003
18. Trueman P, Drummond M, Hutton J: Developing guidance for budget impact analysis. Pharmacoeconomics 19:609-621, 2001
19. Drummond MF, Sculpher MJ, Torrance GW, et al: Methods for the Economic Evaluation of Health Care Programmes, 3rd ed. New York: Oxford University Press; 2015
20. Sanders GD, Neumann PJ, Basu A, et al: Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. JAMA 316:1093-1103, 2016
21. Shih YC, Halpern MT: Economic evaluations of medical care interventions for cancer patients: how, why, and what does it mean? CA Cancer J Clin 58:231-244, 2008
22. Shaya FT, Breunig IM, Seal B, et al: Comparative and cost effectiveness of treatment modalities for hepatocellular carcinoma in SEER-Medicare. Pharmacoeconomics 32:63-74, 2014
23. Stinnett AA, Mullahy J: Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Mak 18: S68-S80, 1998
24. Tambour M, Zethraeus N, Johannesson M: A note on confidence intervals in cost-effectiveness analysis. Int J Technol Assess Health Care 14:467-471, 1998
25. Hoch JS, Briggs AH, Willan AR: Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. Health Econ 11:415-430, 2002
26. Shih YC, Pan IW, Tsai YW: Information technology facilitates cost-effectiveness analysis in developing countries: an observational study of breast cancer chemotherapy in Taiwan. Pharmacoeconomics 27:947-961, 2009
27. Goto D, Shih YT, Lecomte P, et al: Regression-based approaches to patient-centered cost-effectiveness analysis. Pharmacoeconomics 35:689-695, 2017