Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2016 April ; 25(4): 462–466. doi:10.1002/pds.3872.

More realistic power estimation for new user, active comparator studies: an empirical example

Mugdha Gokhale, MS[1], **John B. Buse, MD, PhD**[2], **Virginia Pate, MS**[1], **M. Alison Marquis, MSTAT**[3], and **Til Stürmer, MD, PhD**[1]

[1]Department of Epidemiology, University of North Carolina at Chapel Hill

[2]Department of Medicine, University of North Carolina School of Medicine at Chapel Hill

[3]Collaborative Studies Coordinating Center, Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

Purpose—Pharmacoepidemiologic studies are often expected to be sufficiently powered to study rare outcomes, but there is sequential loss of power with implementation of study design options minimizing bias. We illustrate this using a study comparing pancreatic cancer incidence after initiating dipeptidyl-peptidase-4 inhibitors (DPP-4i) versus thiazolidinediones or sulfonylureas.

Methods—We identified Medicare beneficiaries with at least one claim of DPP-4i or comparators during 2007–2009 and then applied the following steps:1) Exclude prevalent users 2) Require a second prescription of same drug 3) Exclude prevalent cancers 4) Exclude patients age <66 years and 5) Censor for treatment changes during follow-up. Power to detect hazard ratios (effect measure strongly driven by the number of events) >=2.0 estimated after step 5 was compared with the naïve power estimated prior to step 1.

Corresponding Author and requests for reprints: Mugdha Gokhale, MS, Department of Epidemiology, University of North Carolina at Chapel Hill, Campus Box 7435, 2106 McGavran-Greenberg Hall, Phone number: 919-904-2637, Fax number : 919-966-6025, mgokhale@unc.edu.

Prior presentations – This research was presented in poster format at ICPE 2014 in Taipei.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Disclosures

M.G. is a doctoral student at UNC Chapel Hill. V.P. receives salary support from investigator initiated grants from Merck and Amgen. J. B. is an investigator and/or consultant without any direct financial benefit under contracts between his employer and the following companies: Andromeda, Astellas, AstraZeneca, Boehringer-Ingelheim, Bristol-Myers Squibb, Dance Pharmaceuticals, Elcylex, Eli Lilly, GI Dynamics, GlaxoSmithKline, Halozyme, Hoffman-LaRoche, Intarcia Therapeutics, Johnson & Johnson, Lexicon, LipoScience, Medtronic, Merck, Metabolic Solutions Development Company, Metabolon, Metavention, Novartis, Novo Nordisk, Orexigen, Osiris, Pfizer, Rhythm, Sanofi, Takeda, Tolerex, Transtech Pharma, Veritas, and Verva. JBB is a consultant for and will receive stock options from PhaseBio. TS receives investigator-initiated research funding and support as Principal Investigator (R01 AG023178) and Co-Investigator (R01 AG042845) from the National Institute on Aging (NIA), and as Co-Investigator (R01 CA174453) from the National Cancer Institute (NCI) at the National Institutes of Health (NIH), and as Principal Investigator of a Pilot Project from the Patient Centered Outcomes Research Institute (PCORI). He also received research funding as Principal Investigator of the UNC-DEcIDE center from the Agency for Healthcare Research and Quality. Dr. Stürmer does not accept personal compensation of any kind from any pharmaceutical company, though he receives salary support from the Center for Pharmacoepidemiology (current members: GlaxoSmithKline, UCB BioSciences, Merck) and research support from pharmaceutical companies (Amgen, Genentech, Merck, Sanofi) to the Department of Epidemiology, University of North Carolina at Chapel Hill. MM receives salary support from Medtronic.

Results—There were 19,388 and 28,846 DPP-4i and thiazolidinedione initiators during 2007–2009. The number of drug initiators dropped most after requiring a second prescription, outcomes dropped most after excluding patients with prevalent cancer, and person-time dropped most after requiring a second prescription and as-treated censoring. The naïve power (>99%) was considerably higher than the power obtained after the final step (~75%).

Conclusions—In designing new-user active-comparator studies, one should be mindful how steps minimizing bias affect sample-size, number of outcomes, and person-time. While actual numbers will depend on specific settings, application of generic losses in percentages will improve estimates of power compared with the naive approach mostly ignoring steps taken to increase validity.

Keywords

New-user design; Power; Sample size; Bias

Introduction

Studies using large databases often provide large sample sizes[1] and are therefore expected to be sufficiently powered to investigate rare outcomes. Power calculations included in study protocols based on the number of drug claims and outcomes in the database over a fixed time often do not take into account steps taken to increase the internal validity and therefore overestimate power considerably. Accounting for steps taken to minimize the potential for bias will yield more realistic power estimates, but this implies implementing most steps of the study, at which point power calculations are moot.

Power/precision is important while rating the quality of evidence from studies used for systematic reviews or clinical guidelines. The GRADE guidelines, for example, rate the evidence from under-powered observational studies as ‘very low quality’ and recommend rating down the quality of evidence by two levels in the presence of very few events and confidence intervals including both appreciable benefit and harm.[2]

We recently conducted a study comparing pancreatic cancer incidence with dipeptidyl-peptidase-4 inhibitors (DPP-4i) versus sulfonylureas (SU) and thiazolidinediones (TZD) using Medicare claims data from 2007–2011.[3] Preliminary power calculations suggested adequate DPP-4i prescriptions and pancreatic cancers in our datasets, but without data on treatment durations or censoring we were unable to estimate the mean follow-up, which led to considerable overestimation of the number of outcomes. This study demonstrates the loss of new-users, outcomes and person-time after each step taken to minimize bias. We first examined this for Medicare claims data from 2007–2009 and then validated these results with 2010–2012 data. We also report power calculations to detect a clinically meaningful increased hazard ratio of pancreatic cancer based on empirical estimates of follow-up time and other usual parameters.

Methods

This study compared two new-user cohorts DPP-4i versus TZD and DPP-4i versus SU using a 20% random sample of Medicare beneficiaries aged >=66 years with fee-for-service Part A, B, D enrollment in at least one month during a calendar year from January 1, 2007 to December 31, 2011. Medicare covers >98% of US adults >=65 years and contains demographic, medical and pharmacy information for enrollees.[4, 5]

From this, we identified patients with at least one claim of DPP-4i or TZD/SU during 2007–2009 and narrowed down to the final study cohort by excluding 1)Prevalent users of DPP-4i or comparator in the 6 months pre-initiation, 2)Patients without a second prescription of the same drug within 180 days post-initiation, 3)Patients with prevalent cancers, 4)Patients <66 years and 5)Censor for treatment changes during follow-up. We aimed to examine whether pancreatic cancer incidence was higher among DPP-4i initiators relative to comparators, as assessed by the hazard ratio (HR). For HR, power depends strongly on the number of events, a function of the total number of new-users and person-time in each treatment cohort. This study reports the stepwise loss in the proportion of new-users, outcomes and person-time and compares power to detect literature based[6] and clinically meaningful HR (>=2.0) before step 1 and after step 5 calculated using SAS 9.3.[7,8] This process was validated with data from 2010–2012. We also applied empirical estimates of loss in percentages from 2007–2009 data to estimate sample size, person-time and outcomes in the 2010–2012 data and compared the power calculated using empirical estimates with the actual power obtained using 2010–2012 data.

Results

During 2007–2009, there were 19,388 and 28,846 DPP-4i and TZD initiators respectively contributing 50,377 and 78,858 person-years. The mean age in the DPP-4i and TZD groups was 75.1 and 73.5 years. A naïve power calculation at this stage yielded a power of >99% to detect HR >=2.0.

The biggest drop in the proportion of new-users was after restricting to those with a second prescription of the same drug (~20 percentage points; table 1). Excluding prevalent cancers before the second prescription (~29 percentage points) was responsible for the biggest drop in outcomes. Major drops in person-time were due to the second prescription and as-treated censoring. The patterns of drop in percentages during 2010–2012 were very similar (table 2). The power to detect a HR of 2.0 under as-treated analysis was 77% for 2007–2009 and 72% for 2010–2012 which is considerably less than the naïve estimate at step 1. The power obtained with 2010–2012 data after applying the empirical estimates of loss in percentages from 2007–2009 data was 76% which is very close to the actual power obtained with the 2010–2012 data (72%). Similar patterns were observed for the DPP versus SU (supplementary tables 1&2).

Discussion

We illustrate the stepwise loss of drug initiators, outcomes and person-time while implementing a new-user active-comparator study of pancreatic cancer incidence with

DPP-4i versus comparators. Major losses in the proportion of new-users and person-time occurred when the population was restricted to those having a second prescription of the drug, a criterion that increases the chance that the patients are actually started on the drug. The biggest drop in outcomes occurred after excluding individuals with evidence of prevalent cancers before the start of follow-up at the 2nd prescription. While we could have only excluded patients with pancreatic cancer at baseline (before the first prescription), cancer diagnoses/treatments between baseline and the start of follow-up might affect the risk for pancreatic cancer or the sensitivity and specificity of its diagnosis and therefore applied this criterion. Finally, the as-treated analysis reduced the outcomes and person-time leading to power estimates considerably less than estimated before step 1.

Statistical power of a study using secondary data is often overestimated at the study protocol stage based on easy-to-get pilot data on number of prescriptions and outcomes. While empirical data is preferred to estimate realistic power, this requires access to data and implementation of close to 100% of study steps, at which point power calculations are moot. Therefore, compared to naïve power calculations, our numbers at the protocol stage combined with rough estimates of number of prescriptions and incidence rates for the outcome in the population of interest should provide an approximation for studies that implement the steps of a new-user design in settings similar to ours. This is of particular relevance to studies comparing effects of diabetes treatments on incident cancers, a major focus of research.[9–11] This does not imply a ‘one-size fits all’ approach to calculate power, however, and our method of comparing two survival curves will clearly not be generalizable to all other studies. Power calculations should always be based on the analysis methods selected to answer the question of interest. If relevant data can be obtained to implement more realistic power calculations, such data should obviously be used rather than our approximations.

Researchers should further be cognizant of additional decreases in power as a result of subgroup analyses, restriction and confounding control methods. Example, there could be substantial losses in the size of the comparator cohort due propensity score matching or in both treatment cohorts due to trimming the tails of propensity score distributions to reduce the potential for unmeasured confounding by frailty.[12,13]

Our study implies that researchers should be alert to the losses in available sample sizes and person-time with efforts to minimize bias. Similar logic would apply to often preferred measures related to the precision of the effect estimates (width or upper /lower limits of the confidence interval).[14] We here focus on power solely because it is most often used at the study protocol stage without implying that we would encourage using statistical testing over reporting of point estimates and their precision.

A strength of our study was that estimates of treatment duration were based on our previously published study. A short treatment duration may be insufficient to observe a causal effect of a drug on cancer incidence because of the induction and latent periods.[15] Note that this issue is related to but not identical to person-time. While person-time can be increased by increasing the cohort size, the treatment duration will be mainly driven by actual treatment dynamics (in addition to restrictions of the database) and thus cannot be

addressed by increasing N. Accounting for empirical estimates of duration of treatment gave realistic power estimates and also allowed us to evaluate whether we had enough patients treated long enough to allow for an induction and latent period.

Our study had some caveats. First, we used an older Medicare population which has high diabetes prevalence and higher incidence of cancers compared to the general population. Therefore our numbers may not be directly applicable to other databases. However, using our numbers at the proposal stage combined with rough estimates of context-specific data on exposure and outcomes will provide a good alternative, compared to naïve power calculations. Second, the follow-up time used was the median follow-up in the DPP-4i group.[3] Using the 75th percentile of follow-up time somewhat increased the power (~90%), but was still lower than the naïve estimates.

In summary, while designing new-user active-comparator studies, one should be mindful how steps to minimize the potential for bias affect sample size, outcomes and person-time at risk. Our results may be particularly helpful to derive realistic estimates of power (or precision) while designing robust studies comparing rare outcomes for antihyperglycemic drugs or other settings with additional context-specific data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding - Funding for this project was provided by AstraZeneca (Wilmington, Delaware). The project described was also supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number 1UL1TR001111 and R01 AG023178 from the National Institute on Aging. The database infrastructure used for this project was funded by the Pharmacoepidemiology Gillings Innovation Lab (PEGIL) for the Population Based Evaluation of Drug Benefits and Harms in Older US Adults (GIL 200811.0010), the Center for Pharmacoepidemiology, Department of Epidemiology, UNC Gillings School of Global Public Health; the CER Strategic Initiative of UNC's Clinical Translational Science Award (1 ULI RR025747); the Cecil G. Sheps Center for Health Services Research, UNC; and the UNC School of Medicine.

References

1. Strom, BL.; Kimmel, SE.; Hennessy, S., editors. Textbook of Pharmacoepidemiology. 2. Wiley Blackwell; 2013.
2. Guyatt GH, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011; 64:1283–93. [PubMed: 21839614]
3. Gokhale M, Buse JB, Gray CL, Pate V, Marquis MA, Stürmer T. Dipeptidyl – peptidase – 4 inhibitors and pancreatic cancer: A cohort study. Diabetes, Obesity and Metabolism. 2014
4. [Accessed April, 2014] Brief summaries of medicare & medicaid. 2011. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareProgramRatesStats/downloads/MedicareMedicaidSummaries2011.pdf
5. [Accessed April, 2014] Medicare program - general information. http://www.cms.gov/Medicare/Medicare-General-Information/MedicareGenInfo/index.html
6. Elashoff M, Matveyenko AV, Gier B, Elashoff R, Butler PC. Pancreatitis, pancreatic, and thyroid cancer with glucagon-like peptide-1–based therapies. Gastroenterology. 2011; 141(1):150–156. [PubMed: 21334333]
7. Hobbs, G. [Accessed 12/9, 2014] Power analysis: What is available and what you need to know. http://www.wuss.org/proceedings11/Papers_Hobbs_G_76155.pdf

8. [Accessed 12/9, 2014] TWOSAMPLESURVIVAL statement. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_power_sect014.htm
9. Nguyen QT, Sanders L, Michael AP, Anderson SR, Nguyen LD, Johnson ZA. Diabetes medications and cancer risk: Review of the literature. American Health & Drug Benefits. 2012; 5(4)
10. Bolen S, Wilson L, Vassy J, et al. Comparative effectiveness and safety of oral diabetes medications for adults with type 2 diabetes. 2007
11. Drzewoski J, Drozdowska A, Sliwinska A. Do we have enough data to confirm the link between antidiabetic drug use and cancer development. Pol Arch Med Wewn. 2011; 121(3):81–87. [PubMed: 21430609]
12. Glynn RJ, Schneeweiss S, Stürmer T. Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology. Basic & clinical pharmacology & toxicology. 2006; 98(3):253–259.10.1111/j.1742-7843.2006.pto_293.x [PubMed: 16611199]
13. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment Effects in the Presence of Unmeasured Confounding: Dealing With Observations in the Tails of the Propensity Score Distribution—A Simulation Study. American Journal of Epidemiology. 2010; 172(7):843–854.10.1093/aje/kwq198 [PubMed: 20716704]
14. Bland JM. The tyranny of power: Is there a better way to calculate sample size? BMJ. 2009; 339:1133–5.
15. Rothman KJ. Induction and latent periods. Am J Epidemiol. 1981 Aug; 114(2):253–9. [PubMed: 7304560]

Example SAS code for power calculations

```
*DPP vs TZD 2007 – 2009;
proc power;
twosamplesurvival test=logrank
      curve("unexposed")= 1:0.998 2:0.996 3:0.995 /* based on the kaplan
meier curve. time:survival at various points in the TZD group*/
      refsurvival = "unexposed"
      hazardratio = 2.0 /*power to detect hazard ratio of 2.0 */
       accrualtime = 0.01 /* time during which subjects are recruited in
the study. Set to a minimal non-zero value 0.01 since we are not
recruiting/accruing any subjects*/
      followuptime = 0.8 /*median followup time in the data */
      groupns = 19388|28846 /* number of new users in TZD and DPP groups
respectively*/
      power = .;
run;
```

Take home messages

- In pharmacoepidemiologic studies, the power or sample size estimates reported in the study protocol based on easy-to-get pilot data are often naïve because they ignore the study design steps implemented to minimize the potential for bias.
- This study illustrates the stepwise loss of drug initiators, outcomes, person-time while implementing a new-user active-comparator cohort study of pancreatic cancer incidence with dipeptidyl-peptidase-4 inhibitors (DPP-4i) versus thiazolidinediones (TZD) or sulfonylureas (SU).
- In a population of Medicare enrollees >= 66 years of age, the biggest drop in sample size occurred with requiring a second prescription of the same drug; the biggest drop in number of outcomes occurred after excluding patients with prevalent cancer; the biggest drops in person-time occurred after requiring a second prescription and as-treated censoring for treatment changes. The naïve statistical power to detect hazard ratio >= 2.0 was considerably higher (>99%) than the power obtained at the final step (~75%) for both cohorts.
- Our results are of particular relevance for studies comparing the effect of diabetes treatments on incident cancers. While actual numbers will depend on the specific setting, application of generic percentages of loss in sample size and person-time will improve power estimates in other studies comparing outcomes for antihyperglycemic drugs and, most likely, also other studies with context-specific data compared with the naïve approach ignoring the effects of study design to improve validity on power and sample size.

Table 1

Number of patients and pancreatic cancer outcomes for dipeptidyl-peptidase-4 inhibitors (DPP4i) and thiazolidinediones (TZD) groups : 2007–2009

	DPP-4i			TZD		
	N	**Pancreatic cancer Outcomes**	**Total person- time (years)**	**N**	**Pancreatic cancer Outcomes**	**Total person- time (years)**
Number of DPP or TZD pharmacy claims 2007 – 2009	537937	NA	NA	2439040	NA	NA
Patients with at least 1 script of DPP or TZD between 2007 – 2009	31868	NA	NA	32044	NA	NA
Step 1: New users without use of DPP or TZD during 6 months washout[#]	19388 (100%)	74 (100%)	50377 (100%)	28846 (100%)	93 (100%)	78858 (100%)
Step 2: Restricting to new-users with a second script within 180 days of the index date[^]	15893 (82.0%)	71 (95.9%)	40256 (79.9%)	23068 (79.9%)	88 (94.81%)	61510 (78.0%)
Step 3: Excluding patients with prevalent cancers[^]	13005 (67.1%)	49 (66.2%)	37956 (75.3%)	19706 (68.3%)	62 (66.7%)	58785 (74.5%)
Step 4: Excluding age <66 (final study cohort)[^]	12208 (63.04%)	47 (63.5%)	35496 (70.4%)	16672 (57.8%)	55 (59.1%)	49259 (62.4%)
Step 5: As-treated analysis [*]	12208 (63.04%)	35 (47.2%)	22572 (44.8%)	16672 (57.8%)	38 40.9%)	26689 (33.8%)

[#]For step 1, person-time calculated from the date of initiation till the earliest of outcome occurrence, death, end of enrollment or end of study.

[^]For steps 2,3,4, person-time calculated from the date of the second prescription till the earliest of outcome occurrence, death, end of enrollment or end of study.

[*]As-treated analysis – patients were followed from the second prescription to the earliest of the following – outcome occurrence, treatment switching/discontinuation/augmentation, death, end of enrollment or end of study.

The naïve power before step 1 was estimated using the ratio of TZD to DPP-4i initiators, reference risk of outcome, and assuming a closed cohort with no loss to follow-up. The power after step 5 on the final cohort was calculated using the “twosamplesurvival” procedure in SAS 9.3, assuming an empirical median treatment duration of 9 months.[3]

Table 2

Number of patients and pancreatic cancer outcomes for dipeptidyl-peptidase-4 inhibitors (DPP4i) and thiazolidinediones (TZD) groups : 2010–2012

	DPP-4i			TZD		
	N	**Pancreatic cancer Outcomes**	**Total person-time (years)**	**N**	**Pancreatic cancer Outcomes**	**Total person- time (years)**
Number of DPP or TZD pharmacy claims 2011 – 2012	1333032	NA	NA	1369955	NA	NA
Patients with at least 1 script of DPP or TZD between 2011 – 2012	70118	NA	NA	26251	NA	NA
Step 1: New users without use of DPP or TZD during 6 months washout[#]	46097 (100%)	95 (100%)	50014 (100%)	20085 (100%)	64 (100%)	31923 (100%)
Step 2: Restricting to new-users with a second script within 180 days of the index date[^]	36055 (78.2%)	94 (98.9%)	36853 (73.7%)	15144 (75.4%)	62 (96.9%)	22990 (72.0%)
Step 3: Excluding patients with prevalent cancers[^]	29525 (64.0%)	52 (54.7%)	34150 (68.2%)	12991 (64.7%)	46 (71.9%)	21814 (68.3%)
Step 4: Excluding age < 66 (final study cohort)[^]	27303 (59.2%)	49 (51.6%)	31425 (62.3%)	10216 (51.41%)	39 (60.9%)	17280 (54.1%)
Step 5: As-treated analysis [*]	27303 (59.2%)	39 (41.1%)	25569 (51.1%)	10216 (51.41%)	20 (31.3%)	10430 (32.6%)

[#]For step 1, person-time calculated from the date of initiation till the earliest of outcome occurrence, death, end of enrollment or end of study.

[^]For steps 2,3,4, person-time calculated from the date of the second prescription till the earliest of outcome occurrence, death, end of enrollment or end of study.

[*]As-treated analysis – patients were followed from the second prescription to the earliest of the following – outcome occurrence, treatment switching/discontinuation/augmentation, death, end of enrollment or end of study.

The naïve power before step 1 was estimated using the ratio of TZD to DPP-4i initiators, reference risk of outcome, and assuming a closed cohort with no loss to follow-up. The power after step 5 on the final cohort was calculated using the “twosamplesurvival” procedure in SAS 9.3, assuming an empirical median treatment duration of 9 months.[3]