EDE14-324

Brief Report

The quasi-cohort approach in pharmacoepidemiology: Upgrading the nested case-control

Samy Suissa

Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada
McGill Pharmacoepidemiology Research Unit, Centre for Clinical Epidemiology, Jewish General Hospital, Montreal, Canada.

Please address correspondence and request for reprints to:

Samy Suissa
Centre for Clinical Epidemiology, Jewish General Hospital
3755 Cote Ste-Catherine, H4.61
Montreal, Québec, Canada H3T 1E2
Tel: 514-340-7593 Fax: 514-340-7564
E-mail: samy.suissa@mcgill.ca

Words text: 1,500 (Brief Report) **Words abstract: 150**

Running title: The quasi-cohort approach

FUNDING SOURCE

This research was funded in part by a grant from the Canadian Institutes of Health Research (CIHR) and the Canadian Foundation for Innovation (CFI). The author is the recipient of the James McGill Chair award.

ACKNOWLEDGEMENTS

I thank the *Commission d'accès à l'information du Québec* and the *Régie de l'assurance maladie du Québec* (RAMQ) for the database used in the illustration.

Date submitted: 05 May 2014
Date accepted: 19 July 2014

No. of Text Pages 11
No. of Tables 4
No. of Figures 0

July 15, 2014

ABSTRACT

Observational studies of drug effects conducted using health care mega-databases often involve large cohorts with multiple time-varying exposures and covariates. These present formidable technical challenges in data analysis, necessitating sampling approaches such as nested case-control designs. The nested case-control approaches, however, baffling to medical journal readers, particularly the comparisons involving "cases" versus "controls" and the convoluted way in which forward-looking relations from exposure to outcome are extracted from backward-looking data. I propose a "quasi-cohort" approach involving alternative ways of data presentation and analysis that are more in line with the underlying cohort design, including the computation of quasi-rates, rate ratios and quasi-rate differences. I illustrate the quasi-cohort approach using data from a study of pneumonia risk associated with inhaled corticosteroid use in a cohort of 163,514 patients with chronic obstructive pulmonary disease, including 20,344 who had the outcome event of pneumonia hospitalization during more than 304 million person-days of follow-up.

Observational studies conducted using existing huge health care databases have become the standard in assessing the effects of drugs. These studies typically involve large cohorts in which, often, the drug exposure under study and confounding factors vary over time. These variables thus need to be recomputed at every new time point of follow-up, which implies complex measures of exposure and formidable technical challenges in data analysis. For example, a recent study of the effect of antihypertensive drugs on the risk of cancer involved a cohort of 1,165,781 patients followed for up to 14 years, for a total density of over 2.7 billion patient-days.[1] Consequently, the analysis of the entire cohort becomes impossible, and designs such as nested case-control, based on sampling from the cohort, have instead been used.[2,3] This approach, first called a synthetic retrospective study, was subsequently developed as "case-control within a cohort." [4-6]

Several misconceptions regarding the nested case-control design endure among editors and reviewers of medical journals (which is where an increasing number of such observational studies are published). Indeed, the concept of selecting "controls" from a cohort, designed to estimate a rate ratio, is often misunderstood as a selection of persons, rather than person-moments, with resulting confusion when the number of controls exceeds the number of subjects in the cohort. As well, the presentation of the resulting data as "cases" versus "controls" can confuse many reviewers and readers alike, particularly as the cases will systematically be "sicker" than the controls. Finally, as the natural scientific chronology is forward-looking from exposure to outcome, the unnatural direction of the case-control approach from outcome back to exposure creates challenges in recognizing the resulting effect measures as forward-looking.

Major culprits in these misunderstandings are in the data analysis and data presentation, as well as in the "case-control" label itself – referring to a design unfairly seen as inferior compared with cohort studies, even if it simply represents an analysis strategy of the cohort.

In this paper, I introduce alternative ways of presenting data from the nested case-control design and propose the label "quasi-cohort," which better reflects the nature and value of the underlying cohort design. I describe the computation of quasi-rates, which are more in line with the familiar cohort approach, and describe modeling techniques to estimate rate ratios and quasi-rate differences. Finally, I illustrate the design using data from a study of the risks of pneumonia associated with the use of inhaled corticosteroids in chronic obstructive pulmonary disease (COPD).

THE QUASI-COHORT APPROACH

The quasi-cohort approach involves selecting all outcome events from a cohort, along with their exposure classification at the moment of the event, and selecting a sample of person-moments from the cohort follow-up, which can be done in several ways. One is a random sample of *n* person-moments from the sample space of all *N* person-moments generated by the cohort follow-up.[7] Alternatively, quasi-cohort person-moments can be selected from the risk set defined by the timing of the outcome event.[8]

Presentation of quasi-cohort data

A common misunderstanding of the nested case-control approach arises from the presentation of covariates, comparing "sicker cases" versus "controls," inherent in the first table of the reports of such studies. Instead, the first table in reports of the proposed quasi-cohort approach is a comparison between exposure categories in the selected quasi-cohort sample. This approach reflects more faithfully the underlying cohort nature of the study and focusses the assessment of imbalances in the confounders on the associations with exposure rather than as risk factors for the outcome. Such a table would thus not draw the typical unwarranted criticism directed to case-control comparisons.

Second, in the nested case-control approach, the tables presenting effects of drug exposure are also displayed as a comparison of drug exposure prevalence among the cases and controls. Such data are also difficult to grasp for the casual clinical journal reader, who is looking for the effect of exposure on outcome, but presented with data in the opposite direction, namely the "effect" of outcome on prior exposure. Rather, the quasi-cohort approach proposes to present "quasi-rates" for each exposure as well as the corresponding estimated rate ratios. As shown in Table 1, quasi-rates are computed as $(x_i/n_i)(n/N)$, namely the "rates" from the quasi-cohort multiplied by the sampling fraction, with corresponding rate ratios.

Estimation of adjusted rate differences

An important alternative measure of effect is the rate difference, which provides a measure of the impact of the drug exposure in absolute, rather than relative, terms.[9] Many journals now require studies, including case-control studies, to include such an additional

measure of impact. Table 1 and the Appendix describe two such methods of estimating the rate difference.

ILLUSTRATION

To illustrate the quasi-cohort approach, we use a cohort of patients with COPD formed from the health insurance databases of the province of Quebec, Canada. This cohort includes 163,514 patients newly treated during 1990-2005 and followed through 2007, with 20,344 who had the outcome event of hospitalization for pneumonia during the 5.4 years of follow-up (overall incidence rate 24.4/1000/year). The study question is whether inhaled corticosteroids increase the risk of serious pneumonia. Since the relevant risk under study is suspected to occur only under current use and disappear once exposure is halted, and given that inhaled corticosteroids are often used irregularly, it is crucial to measure exposure on a daily basis, making the day the time-unit of analysis. Since the cohort generates an incidence density of 304,646,593 person-days of follow-up and involves several time-varying variables, a quasi-cohort approach is inevitable. For the purpose of the illustration, we selected a 4-fold quasi-cohort (size four times the number of outcome events) as a random sample of 81,376 person-moments from the cohort density, as well 1-, 10- and 100-fold sizes.

Table 2 describes the potential confounding factors contrasted by the three exposure categories under consideration from the 4-fold quasi-cohort selected by incidence density random sampling from the over 304 million person-days of follow-up generated by the cohort. Current use is defined as use at the time of the selected person-

moment; no use is defined by no prescriptions for inhaled corticosteroids in the year prior to the selected person-moment; and discontinued use refers to use that stopped over 60 days prior to the selected person-moment.

Table 3 displays the numbers of events and quasi-person-moments, as well as the corresponding quasi-rates and rate ratios for current and discontinued inhaled corticosteroids use relative to no use using the different sized quasi-cohorts.

Table 4 shows that, using the overall rate of pneumonia hospitalization of 2.24 per 1000 per year in the entire cohort, the adjusted rate ratio of 2.28 for current inhaled corticosteroids use can be converted to an approximate adjusted rate difference of 23.5 (95% confidence interval [CI]= 22.5-24.5) additional pneumonia hospitalizations per 1000 per year with current use of inhaled corticosteroids. Alternatively, it also shows that, using the sampling fraction of 81,376 over 304.6 million, the additive odds model produces a rate difference estimate of 19.6 (95% CI= 18.5-20.7) additional pneumonia hospitalizations per 1000 per year with current use of inhaled corticosteroids.

DISCUSSION

Cohort studies conducted within existing computerized health care mega-databases can be so large that they are technically unmanageable for data analysis and thus sampling designs within the cohort become unavoidable. In this paper, we propose to call such designs "quasi-cohort," rather than the common "nested case-control" label that has led to misunderstanding in specialty journals. We also provide formulae and models to analyse the data in ways more in line with cohort studies, using quasi-rates and quasi-rate

differences, resulting in presentation of the data that is in unison with the underlying cohort.

The changes proposed in this paper stem from some misconceptions regarding the nested case-control design. Indeed, the selection of "controls" from a cohort is generally misunderstood as a selection of persons, not person-moments, leading to confusion when the number of controls exceeds the number of subjects in the cohort (such as a cohort of 163,514 patients from which 197,705 "controls" were selected[10]). Other sources of confusion include the presentation of data as a comparison between "cases" and "controls," as well as the convoluted way that forward-looking associations from exposure to outcome are extracted from backward-looking data. The quasi-cohort approach eliminates these concerns.

Sampling of person-moments is not always necessary, such as when estimating the cumulative incidence ratio, where persons can be sampled by the nested case-control design. In this case, however, the full cohort analysis should not pose any technical issue.

I also addressed the growing demand for absolute measures of excess risk, such as the rate difference, in medical journals.[8] I provided two approaches to estimate adjusted rate differences, though more theoretical work on these approaches is still needed.

In summary, this paper proposes the label "quasi-cohort" rather than "nested case-control" to designate study designs and data analyses based on sampling within cohorts as a more accurate reflection of the underlying cohort and intent of the strategy. With the computation of quasi-rates and corresponding rate ratios, this approach should facilitate the review of the many studies that use such sampling schemes within mega-cohorts,

particularly with the proposed alternative way of presenting data from the quasi-cohort approach and the tools provided to estimate excess risk measures.

References

(1) Azoulay L, Assimes TL, Yin H, Bartels DB, Schiffrin EL, Suissa S. Long-term use of angiotensin receptor blockers and the risk of cancer. *PLoS One.* 2012;7(12):e50893.

(2) Suissa S. Novel approaches to pharmacoepidemiological study design and statistical analysis. In: Strom B, ed. *Pharmacoepidemiology*. 4th edition ed. New York: John Wiley's & Sons, 2005: 811-30.

(3) Essebag V, Genest J, Jr., Suissa S, Pilote L. The nested case-control study in cardiology. *Am Heart J.* 2003;146(4):581-590.

(4) Mantel N. Synthetic retrospective studies and related topics. *Biometrics.* 1973;29(3):479-486.

(5) Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis appraisal by application to asbestos mining. *J R Stat Soc A.* 1977;140:469-491.

(6) Thomas DC. Addendum to paper by Liddell FDK, McDonald JC and Thomas DC. *J R Stat Soc A.* 1977;140:483-485.

(7) Miettinen OS. Estimability and estimation in case-referent studies. *Am J Epidemiol.* 1976;103(2):226-235.

(8) Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of Controls in Case-Control Studies: III. Design Options. *Am J Epidemiol.* 1992;135 (9):1042-1050.

(9) Gail MH. Using absolute risks to assess the risks and benefits of treatment. *Thorax.* 2014.

(10) Suissa S, Patenaude V, Lapi F, Ernst P. Inhaled corticosteroids in COPD and the risk of serious pneumonia. *Thorax.* 2013;68(11):1029-1036.

APPENDIX (words: 343)

Estimation of quasi-rate differences

An important alternative measure of the effect of drug exposure on the outcome is the excess risk measured by the rate difference, which provides a measure of the impact of the drug exposure in absolute rather than relative terms.[8] Many journals in fact require studies, including case-control studies, to include such a calculation as an additional measure of impact.

Table 1 provides the estimator of the crude quasi-rate difference, obtained directly from the quasi-rates. To estimate the adjusted quasi-rate difference, one can use the adjusted rate ratio (RR) estimated by the logistic regression model, after adjustment for covariates, along with the overall rate of the outcome event (R_t) from the full cohort simply computed from the known total cohort person-time. The resulting adjusted quasi-rate difference (RD) for a dichotomous exposure can then be approximated by

$$RD = R_t\,(RR\text{-}1) \,/\, (P_0 + P_1\,RR)$$

where P_1 and P_0 denote the prevalence of exposed and unexposed respectively (P_1+P_0=1) estimated from the selected quasi-cohort person-moments. This formula can be generalized if the exposure is polytomous and if the desired rate difference is between one of the several exposure categories and a reference category to

$$RD = R_t\,(RR_1\text{-}1) \,/\, (P_0 + \sum P_k\,RR_k)$$

where RR_k is the estimated rate ratio for exposure category k relative to the reference (k=1 to c), P_k and P_0 denote the prevalence of exposure for the different categories and the reference respectively ($P_0 + \sum P_k$ =1), estimated from the quasi-cohort.

The second approach to estimate the adjusted quasi-rate difference is based on directly modeling the quasi-cohort data using a generalized linear additive model for the odds of the outcome event (1=event, 0=quasi-cohort sample), corrected for the sampling fraction. This can be done with a "odds" link function, namely by fitting $R/(1-R)$ as a linear combination of the exposures and covariates, where R is the probability of the outcome event at a person-moment, and using a binomial distribution. The resulting coefficients must then be corrected by the sampling fraction (n/N), to produce the quasi-rate differences.

Table 1

Table 1. Data structure from a full cohort analysis with a dichotomous exposure measured at each of the N person-moments and a quasi-cohort analysis, based on a sample of n person-moments, formed using all outcome events and an incidence density random sample from the cohort, to describe the estimation of the quasi-rates, rate ratio and rate difference

Full cohort analysis

Exposed	Outcome events	Person-moments	Rate of outcome per person-moment	Rate ratio	Rate difference
Yes	x_1	N_1	x_1 / N_1	$(x_1/N_1) / (x_0/N_0)$	$(x_1/N_1) - (x_0/N_0)$
No (reference)	x_0	N_0	x_0 / N_0	1.0	0.0
Total	x	N	x / N		

Quasi-cohort analysis

Exposed	Outcome events	Quasi person-moments	Quasi-rate of outcome per person-moment	Rate ratio	Quasi-rate difference
Yes	x_1	n_1	(n/N) (x_1 / n_1)	$(x_1/n_1) / (x_0/n_0)$	$(n/N)[(x_1/n_1)-(x_0/n_0)]$
No (reference)	x_0	n_0	(n/N) (x_0 / n_0)	1.0	0.0
Total	x	n	(n/N) (x / n)		

[a] Quasi-cohort of size n person-moments selected from all N person-moments of the full cohort follow-up, ie sampling fraction of n/N.

Table 2

Table 2. Characteristics of the quasi-cohort of 81,376 person-moments (four-to-one) selected by incidence density random sampling from the 304.64 million person-days of follow-up generated by the cohort of 163,514 COPD patients identified from the Régie de l'assurance maladie du Québec (RAMQ) databases during 1990-2007, by exposure status to current use of inhaled corticosteroids

	Inhaled corticosteroid use [a]		
	No [b]	*Current*	*Discontinued*
No. person-moments	49,161	17,944	14,271
Age (years); mean (SD)	71.2 (7.8)	70.6 (7.6)	70.[illegible] (7.7)
Male sex (%)	45.5	48.9	41.7
Prior hospitalisation for pneumonia; %	2.3	[illegible].5	2.3
Medication use in the year prior to cohort entry			
No. prescriptions for respiratory drugs; mean	2.0	[illegible]0	2.0
Oral corticosteroids/antibiotics; %	61.7	65.7	68.4
Cardiovascular drugs; %	[illegible]	63.8	65.5
Anti-diabetic agents; %	11.4	9.2	10.9
Antidepressants; %	13.6	14.2	15.1
Central nervous system drugs; %	53.3	49.3	50.1
Osteoporosis drugs; %	5.1	5.9	6.7
NSAIDs; %	36.9	31.8	34.8
Narcotics; %	15.6	15.1	16.9
Anti-rheumatic agents; %	0.8	0.8	0.9

[a] No use refers to no prescriptions of inhaled corticosteroids in year prior to the selected person-moment; current use is defined by a prescription of inhaled corticosteroids in the 60 days prior to the selected person-moment; and discontinued use as some use during the period 60 days to the year prior to the selected person-moment, but not current.
[b] Reference category.

Table 3

Table 3. Quasi-rates and crude and adjusted rate ratios of hospitalization for pneumonia associated with current use of inhaled corticosteroids using various quasi-cohort sizes selected by risk set sampling from the 304.6 million person-days of follow-up generated by the cohort of 163,514 COPD patients identified from the RAMQ databases during 1990-2007

	No. Outcome events	*No. Quasi-cohort person-days*	*Quasi-rates [a] (per 1000 person-years)*	*Crude quasi-rate ratio*	*Adjusted [b] quasi-rate ratio*	*(95% CI)*
Quasi-cohort size: 1-fold						
Number	20,344	20,344				
Inhaled corticosteroid use						
No use [c]	9,453	12,201	18.9	1.00	1.00	
Current use	7,636	4,559	40.9	2.16	2.27	(2.17 - 2.38)
Discontinued use	3,255	3,584	22.2	1.17	1.26	(1.19 - 1.34)
Quasi-cohort size: 4-fold						
Number	20,344	81,376				
Inhaled corticosteroid use						
No use [c]	9,453	49,227	18.7	1.00	1.00	
Current use	7,636	18,082	41.2	2.20	2.28	(2.20 - 2.37)
Discontinued use	3,255	14,127	22.8	1.21	1.27	(1.21 - 1.33)
Quasi-cohort size: 10-fold						
Number	20,344	203,441				
Inhaled corticosteroid use						
No use [c]	9,453	123,735	18.6	1.00	1.00	
Current use	7,636	44,640	41.7	2.24	2.31	(2.24 - 2.39)
Discontinued use	3,255	35,045	22.7	1.22	1.28	(1.23 - 1.33)
Quasi-cohort size: 100-fold						
Number	20,344	2,034,333				
Inhaled corticosteroid use						
No use [c]	9,453	1,232,964	18.7	1.00	1.00	
Current use	7,636	448,340	41.5	2.22	2.26	(2.19 - 2.33)
Discontinued use	3,255	353,029	22.5	1.20	1.26	(1.21 - 1.31)

[a] Quasi-rates computed using person-moments from quasi-cohort and corresponding sampling fraction from the 304.64 million person-days of the full cohort.

[b] Adjusted for factors in Table 2.

[c] Reference category.

Table 4

Table 4. Quasi-rates and crude and adjusted rate differences of hospitalization for pneumonia associated with current use of inhaled corticosteroids using the approximate method and the corrected Poisson regression method for the 4-fold quasi-cohort selected by incidence density random sampling from the 304.6 million person-days of follow-up generated by the cohort of 163,514 COPD patients identified from the RAMQ databases during 1990-2007

	No. Outcome events	*No. Quasi-cohort person-days*	*Quasi-rates* [a] *(per 1000 person-years)*	*Crude quasi-rate differences*	*Adjusted quasi-rate differences*	*(95% CI)*
Approximate multiplicative model						
Number	20,344	81,376				
Inhaled corticosteroid use						
No use [c]	9,453	49,161	18.8	0	0.0	
Current use	7,636	17,944	41.5	22.7	23.5	(22.5 – 24.5)
Discontinued use	3,255	14,271	22.3	3.5	4.6	(3.7 – 5.6)
Corrected additive model						
Number	20,344	81,376				
Inhaled corticosteroid use						
No use [c]	9,453	49,161	18.8	0.0	0.00	
Current use	7,636	17,944	41.5	22.8	19.6	(18.5 – 20.7)
Discontinued use	3,255	14,271	22.3	3.5	3.6	(2.8 – 4.4)

[a] Quasi-rates computed using person-moments from quasi-cohort and corresponding sampling fraction from the 304.64 million person-days of the full cohort.

[b] Adjusted for factors in Table 2.

[c] Reference category.