Epidemiological research based on large data analysis: study characteristics

Omar Hasan Kasule Sr and Lin Naing
Institute of Medicine, Universiti Brunei Darussalam, Tungku Link, Gadong BE1410, Brunei Darussalam

Abstract

Analysis of studies published in one volume of each of 3 major epidemiological studies revealed median study sizes above 1000 for all types of study design, data collection, and study population. Cross sectional and follow up studies had the highest median study sizes when they were based on previously and routinely collected data. The paper discusses some of the problems associated with large studies.

Introduction

This study was motivated by the observation that recent epidemiological research is based on existing large databases or defined cohorts and that 100% sampling was the usual practice. This is a reversal of the traditional epidemiological practice of selecting a small probability sample from a study population in order to reach conclusion about the target population [1].

Developments in information technology and mass access to the internet opened up new fields of endeavor for the epidemiologist. For example data could be collected from a large number of people using internet-based questionnaires [2]

The objective of the research was to survey epidemiological research published in 2006 in three high-impact journals to make a statistical description of the characteristics of these studies. The three journals selected for study were the American Journal of Epidemiology 2006 Volume 169 Nos. 1-12, The International Journal of Epidemiology 2006 Volume 35 Nos. 1-4, and The European Journal of Epidemiology 2006 Volume 21 Nos. 1-9.

Correspondence:
Omar Hasan Kasule
Institute of Medicine,
Universiti Brunei Darussalam,
Tungku Link, Gadong BE1410,
Brunei Darussalam

Methods

The study included only original papers that involved raw data. Reviews, meta-analyses, and analyses based on published data were excluded. A pre-tested data abstract form was used to abstract the following essential information from each original research article: title, authors, issue and volume number, date of publication, type of study (cross sectional, case control, cohort, randomized community control, randomized clinical), Study population (general population, defined population, ongoing study), type of data collection (routinely collected data, newly collected data, previously collected data or a combination among the above) and total number of study subjects (number recruited before any exclusions). Defined populations were hospitals, health insurance of health maintenance organizations, clinics, schools, factories, and ongoing studies. For case control studies cases and controls were added up. For ongoing studies it was assumed that data collection was new unless a special mention was made of using previously collected data. The data was keyed into an SPSS data base for categorical analysis using the chi-square test statistic to test for association. The Kruskall-Wallis non-parametric test statistic was used to compare study size by various study designs and types of data collection.

Results

A total of 137 studies were analyzed. Table 1 shows a significant variation in journal preference for study designs and methods of data collection. There was however no significant variation among journals in the choice between defined and general study populations.

Table 1 Classifications of articles by journal

Study Characteristics		IJE	EJE	AJE	Total	$X^2(df)$	*P* value[a]
		n (%)	*n* (%)	*n* (%)			
Study design	Cross sectional	16 (39.0)	33 (60.0)	13 (31.7)	62	10.10 (4)	0.039
	Case control	9 (22.0)	9 (16.4)	7 (17.1)	25		
	Follow up	16 (39.0)	13 (23.6)	21 (51.2)	50		
Study population	Defined population	12 (29.3)	23 (41.8)	22 (53.7)	57	5.02 (2)	.081
	General population	29 (70.7)	32 (58.2)	19 (46.3)	80		
Data Collection	Newly & routinely	6 (14.6)	5 (9.1)	3 (7.3)	14		0.030[b]
	Newly	13 (31.7)	28 (50.9)	25 (61.0)	66		
	Previously	0 (0.0)	1 (1.8)	3 (7.3)	4		
	Routinely	22 (53.7)	21 (38.2)	10 (24.4)	53		

[a] Chi-square test; [b] Fisher's exact test;
IJE = International Journal of Epidemiology
EJE = European Journal of Epidemiology
AJE = American Journal of Epidemiology

Table 2 shows that the median study side was over 1000 for all journals and types of study design. Its variation among journals was significant for follow up studies and not cross sectional and case control studies. The median study size did not vary significantly among the 3 journals for different study populations and methods of data collection.

Table 2 Number of research subjects by study characteristics and journal

Study characteristics	Journal	*n*	Number of research subjects			$X^2.(df)$	*P* value[a]
			Median	Min.	Max.		
Study design							
Cross sectional	IJE	16	7,183	107	212,467,094	0.63 (2)	.730
	EJE	33	4,599	112	36,000,000		
	AJE	13	2,255	139	212,467,094		
Case control	IJE	9	1,263	288	166,310	1.61 (2)	.446
	EJE	9	1,051	372	2,222,404		
	AJE	7	1,875	730	212,467,094		
Follow up	IJE	16	60,925	1,016	60,000,000	7.80 (2)	.020
	EJE	13	9,778	34	11,000,000		
	AJE	21	2,446	209	212,467,094		
Study population							
Defined population	IJE	12	6,381	726	246,146	2.47 (2)	.291
	EJE	23	1,272	34	6,240,130		
	AJE	22	2,102	299	1,299,177		
General population	IJE	29	14,495	107	212,467,094	1.36 (2)	.506
	EJE	32	7,404	188	36,000,000		
	AJE	19	10,932	139	212,467,094		

Data Collection							
	IJE	6	2,380	274	246,146		
Newly & routinely	EJE	5	11,081	1,051	2,222,404	0.75 (2) 0	.686
	AJE	3	11,234	1,068	212,467,094		
	IJE	13	3,290	288	14,495		
Newly	EJE	28	937	34	6,240,130	3.90 (2)	0.142
	AJE	25	2,010	139	21,610		
	IJE	-	-	-	-		
Previously	EJE	1	56,214	56,214	56,214	0.20 (1) 0	.655
	AJE	3	1,516	619	212,467,094		
	IJE	22	180,155	107	212,467,094		
Routinely	EJE	21	19,801	212	36,000,000	2.43 (2) 0	.296
	AJE	10	399,910	1,832	212,467,094		

[a] Kruskal-Wallis Test
IJE = International Journal of Epidemiology
EJE = European Journal of Epidemiology
AJE = American Journal of Epidemiology

Table 3 shows that cross sectional and follow up studies had significantly higher median study size in general populations than in defined populations. No such significant variation was seen in case control studies.

Table 3 Number of research subjects by study design and study population

Study design S	tudy pop.	*n*	Number of research subjects			*Z*	*P* value[a]
			Median	Min	Max		
Cross sectional	Defined pop.	22	2,222	112	6,240,130	-2.28	0.023
	General pop.	40	10,832	107	212,467,094		
Case control	Defined pop.	10	1,552	726	2,222,404	-1.00	0.318
	General pop.	15	1,083	288	212,467,094		
Follow up	Defined pop.	25	2,311	34	1,299,177	-3.77	<0.001
	General pop.	25	83,875	188	212,467,094		

[a] Mann-Whitney test

Table 4 shows that follow up studies had higher median study size if based on previously and routinely collected data. No such significant variation was observed for cross sectional and case control study designs.

Table 4 Number of research subjects by study design and type of data collection

Study design	Data collection	*n*	Number of research subjects			X^2.(*df*)	*P* value[a]
			Median	Min	Max		
Cross sectional	New & routine	3	7,000	274	11,193	7.49 (3)	0.058
	Newly	33	2,650	112	6,240,130		
	Previously	3	1,516	619	212,467,094		
	Routinely	23	95,000	107	212,467,094		
Case control	New & routine	7	1,678	1,051	212,467,094	2.86 (2) 0	.240
	Newly	13	909	288	4,778		
	Previously	-	-	-	-		
	Routinely	5	1,272	828	166,310		
Follow up	New & routine	40	15,127	11,081	246,146	29.53 (3)	<0.001
	Newly	20	995	34	11,267		
	Previously	1	56,214	56,214	56,214		
	Routinely	25	87,922	212	212,467,094		

[a] Kruskal-Wallis Test

Table 5 shows significant variation of median study size with the type of data collection. Study size in defined populations was highest for newly and routinely collected data whereas in the general population median study size was highest for previously and routinely collected data.

Table 5 Number of research subjects by study population and method of data collection

Study population	Data collection	*n*	Number of research subjects			X^2 (*df*)	*P* value[a]
			Median	Min	Max		
Defined population	New & routine	7	11,234	7,000	2,222,404	12.10 (2)	0.002
	Newly	39	2,010	34	6,240,130		
	Previously	-	-	-	-		
	Routinely	11	1,272	212	1,299,177		
General population	New & routine	7	1,263	274	212,467,094	27.47 (3)	<0.001
	Newly	27	1,653	139	47,859		
	Previously	4	28,865	619	212,467,094		
	Routinely	42	187,530	107	212,467,094		

[a] Kruskal-Wallis Test

Discussion

Median study sizes were highest for cross sectional and follow up studies and when based on previously or routinely collected data. This is due to availability of large data bases with routinely or previously collected information. The availability of large data bases and high speed computers has encouraged epidemiologists to analyze data without probability sampling. A large data set gives very stable parameters but the same degree of precision could have been obtained from a smaller sample. What is lost is the ability of the epidemiologist to inspect a small manageable data set, internalize it, and let his intuition act before the data is analyzed. The more intimate contact of the epidemiologist with the data traditionally accounted for deep understanding and discussion which are missed in the new trend. Easy availability of large databases also encouraged epidemiologists to plunge into data analysis before serious thought about the research questions. In some cases the research questions can be prompted by preliminary analysis which can lead to numerous biases. Use of large data sets has the advantage of external validity which had never been the primary objective of epidemiological research. Epidemiologists have traditionally aimed at carrying out a small study based on probability sampling so that they can easily identify and control confounding and other sources of bias with the ultimate aim of internal validity. They knew that external validity (generalization) would be attained inductively by consideration of several studies that are internally valid. Use of large sets of routinely collected data also raises the issue of the quality of the data which is collected with service and administrative and not research considerations in mind.

References

1. Kasule OH. The transition from sample to population epidemiology. J Uni Malaya Med Center (in press).

2. Ekman A, Dickman PW, Klint A, Wederpass E, Litton JE 2006.. Feasibility of using web-based questionnaires in large population-based epidemiological studies, Eur J Epidemiol 21(2): 103-111.