Journal of Clinical Epidemiology 58 (2005) 246–251

Journal of Clinical Epidemiology

# Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data

M.C. Gulliford[a,*], G. Adams[a], O.C. Ukoumunne[a], R. Latinovic[a], S. Chinn[a], M.J. Campbell[b]

[a]*Department of Public Health Sciences, King's College London, Capital House, 42 Weston St, London SE1 3QD, UK*
[b]*Institute of General Practice and Primary Care, ScHARR, University of Sheffield, Sheffield S5 7AU, UK*

Accepted 30 August 2004

**Abstract**

**Background and Objective:** To describe the association between values for a proportion and the intraclass correlation coefficient (ICC).

**Methods:** Analysis of data obtained from the General Practice Research Database (GPRD) for variation between United Kingdom general practices and results from a Health Technology Assessment (HTA) review for a range of outcomes in community and health services settings.

**Results:** There were 188 ICCs from the GPRD, the median prevalence was 13.1% (interquartile range IQR 3.5 to 28.4%) and median ICC 0.051 (IQR 0.011 to 0.094). There were 136 ICCs from the HTA review, with median prevalence 6.5% (IQR 0.4 to 20.7%) and median ICC 0.006 (IQR 0.0003 to 0.036). There was a linear association of log ICC with log prevalence in both datasets (GPRD, regression coefficient 0.61, 95% confidence interval 0.53 to 0.69, $P < 0.001$; HTA, 0.91, 0.81 to 1.01, $P < 0.001$). When the prevalence was 1% the predicted ICC was 0.008 from the GPRD or 0.002 from the HTA, but when the prevalence was 40% the predicted ICC was 0.075 (GPRD) or 0.046 (HTA).

**Conclusion:** The prevalence of an outcome may be used to make an informed assumption about the magnitude of the intraclass correlation coefficient. 

*Keywords:* Cluster randomization; Cluster sampling; Intraclass correlation coefficient; Kappa; Binary data; Prevalence

## 1. Introduction

Cluster randomized designs are increasingly used to evaluate interventions in community health and in the organization and delivery of medical care. Following the work of Cornfield [1], Donner and Klar [2], and others, it is well known that cluster randomized designs are statistically less efficient than designs in which an equal number of individual subjects are randomized. This is because variation in the outcome between the clusters to be allocated can generally be expected to increase the variance of the intervention effect above that expected from simple random allocation of individual subjects. The sample size requirement for a cluster randomized trial must be inflated using the design effect, or variance inflation factor, which is a function of the average number of individuals sampled per cluster and the intraclass correlation coefficient [3]. The intraclass correlation coefficient (ICC) represents the proportion of the total variance that is accounted for by between-cluster rather than within-cluster variation. The ICC has been estimated from an analysis of variance of the outcome on cluster even when the outcome is binary [4,5].

Sample size calculations must be carried out before data are collected, and these will usually rely on external estimates of the ICC, for which there is an increasing number of sources [3,6,7]. However, estimated ICCs for a given measure may have poor generalizability between different contexts. The alternative is to use results obtained from pilot studies but these will often be imprecise [8]. Consequently, there is a greater degree of uncertainty in sample size estimations for cluster randomized trials than for individually randomized trials. This uncertainty may be reduced to some extent by describing conditions under which the ICC may be expected to vary. In the analysis of binary data, it appears to be well recognized that the ICC may depend on the natural cluster size. Smaller clusters generally show greater degrees of clustering ([2], p. 55). Mickey and Goodwin [9], Katz et al. [10], Katz and Zeger [11], and Slymen and Hovell [12] additionally drew attention to the relationship between the prevalence of an outcome and the design effect. In general, more common outcomes are associated with higher design effects [9]; however, there are few empirical data available to quantify this relationship and offer

* Corresponding author. Tel.: +44-(0)207-848-6631; fax: +44-(0)207-848-6605.
*E-mail address*: martin.gulliford@kcl.ac.uk (M.C. Gulliford).

0895-4356/05/$ – see front matter 
doi: 10.1016/j.jclinepi.2004.08.012

information to researchers who are planning cluster randomized trials.

Our objective was to quantify the relationship between prevalence and the ICC in the context of health services, especially primary health care settings. Here we present data from two different sources: the UK General Practice Research Database, a database which includes data from several hundred United Kingdom general practices; and a previously published Health Technology Assessment review which includes data from a wider range of settings relevant to public health and health services research. We describe and quantify the relationship between the prevalence of the measure and the intraclass correlation coefficient.

## 2. Materials and methods

The General Practice Research Database (GPRD) is a large database that contains clinical records for several hundred general practices in the United Kingdom [13]. The database includes records for more than two million patients between1988 to 2002.

The size and complexity of the GPRD makes it necessary to extract files suitable for analysis in specific studies. The present analyses were based in part on analysis of specially extracted data and also on secondary analysis of files prepared for other studies. The data in the GPRD are subject to quality checks; when a practice's data meet these specifications, the data are said to be 'up to standard' (UTS). For each analysis, we sampled all practices that were contributing UTS data during the study period of interest. We analyzed data for acute as well as chronic conditions, including data for the rate or prevalence of the condition and the prescription of specific classes of drugs in each condition. Each analysis was performed for men and women separately. Details of the data analyzed are given in Table 1. The proposal for the study was approved by the Scientific and Ethical Advisory Group for the GPRD.

For chronic conditions, we analyzed data for variation between practices in the prevalence of coronary heart disease, stroke, insulin-treated diabetes mellitus, and non–insulin-treated diabetes mellitus. These included data from 270 practices contributing UTS data between 1994 and 1998. We used the case definitions as published by the United Kingdom Office for National Statistics [14]. In addition to case definitions, this publication [14] provides detailed data for prevalence of these conditions. For cases diagnosed with coronary heart disease in 1994, we also estimated variation between practices in the proportion of cases who were prescribed each of 15 classes of drugs including thiazide diuretics, loop diuretics, potassium sparing diuretics, potassium sparing and other diuretic combinations, beta-blockers, vasodilators, centrally acting anti-hypertensive drugs, alpha blockers, angiotensin converting enzyme inhibitors, nitrates, calcium channel antagonists, lipid lowering drugs, and glucose monitoring materials, including blood glucose monitoring, urine glucose monitoring, or any glucose monitoring. Drug codes were identified using appropriate categories from the *British National Formulary* [15].

We analyzed data for a cohort of non–insulin-treated diabetic subjects who were first prescribed oral hypoglycemic therapy between 1993 and 1998. These were sampled from all 263 practices that contributed UTS data between 1992 and 1998. We analyzed variation in the prescription of the 15 classes of drugs listed above during the 12 months after oral hypoglycemic drugs were first prescribed. The selection of these subjects is described in detail elsewhere [16].

We also analyzed variations in the rate of consultations among registered subjects for each of fifteen acute conditions listed in Table 1. These analyses were based on data for 646,336 subjects registered with 108 practices in 2000.

Table 1
Description of data analyzed from General Practice Research Database

| Conditions | Year | Practices, no. | Subjects or consultations, total no. | Outcomes |
|---|---|---|---|---|
| Prevalence of chronic diseases (4) | | | | |
| Coronary heart disease | 1994 | 270 | 1,591,362 | Prevalence |
| Stroke | 1994 | 270 | 1,591,362 | Prevalence |
| Insulin-treated diabetes | 1994 | 270 | 1,591,362 | Prevalence |
| Non–insulin-treated diabetes | 1994 | 270 | 1,591,362 | Prevalence |
| Prescribing in coronary heart disease | 1994 | 270 | 52,102 | Proportions prescribed 15 classes of drugs |
| Prescribing in non–insulin-treated diabetes mellitus | 1993–1998 | 262 | 12,222 | Proportions prescribed 15 classes of drugs |
| Acute conditions (15)[a] | 2000 | 108 | 646,336 | Rate of consultations per year |
| Prescribing in acute conditions | 2000 | 108 | 2,062–112,631 | Proportion of consultations with diagnosis and prescribed antibiotics, including penicillins, nonpenicillins or penicillins and nonpenicillins combined |

Analyses were performed for men and women separately.

[a] The 15 acute conditions were: all respiratory infections; chest infection; colds and coryza; cystitis; ear infection; flu; laryngitis; other infections; other respiratory disease; respiratory infections; sinusitis; sore throat; tonsillitis; tracheobronchitis; and upper respiratory tract infections.