STATISTICAL TECHNIQUES TO ADDRESS ERRORS IN MEASUREMENT OF PHYSICAL ACTIVITY IN A CASE-CONTROL STUDY

by

Jie Zhang

A report submitted to the faculty of

The University of Utah

in partial fulfillment of the requirements for the degree of

Master of Statistics

Department of Family and Preventive Medicine

The University of Utah

July 2010

Abstract

Regular physical activity, fitness, and exercise are crucial for the health maintenance and well being of people of all ages. However, increased exercise or work intensity may lead to certain medical conditions. It is conjectured that increased activity intensity may be associated with pelvic floor disorders.

Pelvic floor disorders including urinary leakage and pelvic organ prolapse (dropping of pelvic organs such as the uterus into or out of the vagina) are very prevalent. Limited data suggest that to date physical activity may be beneficial in terms of its general effects and may be harmful to the pelvic floor; however, previous studies examining this association are methodologically flawed. In two case-control studies (collectively termed the PhActs study, [Physical Activity Study]), we will examine the association between life-long physical activity and different disorders of the pelvic floor.

Primary analyses of the PhActs Study will focus on the analyses of strenuous physical activity as an exposure variable for the outcome of stress urinary incontinence (SUI case-control study) and pelvic organ prolapse (POP case-control study).

Secondary analyses will be performed using logistic regression incorporating physical activity measured in different ways, graphical time trends display, subcategories of physical activity and sensitivity analyses to examine and correct for 'errors in measurement' where the data permit. Odds ratios will be adjusted for 'errors in measurements', most likely using regression calibration methods.

The main focus on this MStat project is the analyses which involve adjustment for 'error in measurement'. In this study, existing data from a reproducibility sub-study of women enrolled in the case-control study will be analyzed by statistical techniques, Regression calibration and Simulation/extrapolation (SIMEX), allowing adjustments to be made to produce correct standard errors and test statistics. Regression calibration and simulation/extrapolation will be compared in this secondary study. SAS and STATA will be used in this study.

Regression calibration is a statistical method for adjusting point and interval estimates of effect obtained from regression models commonly used in epidemiology for bias due to measurement error in assessing nutrients or other variables. Regression calibration is appropriate when a gold standard is available in a validation study and a linear measurement error model with constant variance applies *or* when replicate measurements are available in a reliability study and linear

random within-person error can be assumed. The PhActs study includes a sub-study involving replicate measurements of physical activity.

The simulation extrapolation (SIMEX) is a method for addressing measurement error in generalized linear models. This method shares the simplicity of the regression calibration method and is suitable for problems with additive measurement error. SIMEX is a simulation-based method aimed at reducing bias caused by the inclusion of error-prone covariates. Estimates are obtained by adding additional measurement error; a type of re-sampling approach. This re-sampling uncovers the trend of measurement error. Once the trend is estimated, final estimates are obtained by extrapolating back to the case of no measurement error.

This project will compare regression calibration to SIMEX based on data available early in the PhActs study and will make a recommendation on choice of correction, with a view toward the final analysis.

Introduction

Pelvic Floor Disorders

Physical activity is any activity that causes your body to work harder than normal. The actual amount of physical activity you need depends on your health goals and body conditions[1]. Regular physical activity is essential to maintain health and welfare at all ages. However, increasing physical activity intensity or occupational activity may lead to certain medical conditions or adverse effect. Presumably, increased activity may be associated with pelvic floor disorders[2].

Pelvic floor disorders are problems that affect women's pelvic organs -- the uterus (or womb), vagina, bladder, rectum and the muscles that surround and support them. The three primary categories of pelvic floor disorders in American that most women seek treatment for are urinary incontinence (SUI), fecal incontinence and pelvic organ prolapsed (POP) (when one or more of the pelvic organs fall into the vagina)[2,3]. The fecal incontinence condition will not be addressed specifically since for most of the middle-aged women, the cause of fecal incontinence is not necessarily related to the pelvic floor but rather to the other conditions like irritable bowel syndrome[4]. Urinary incontinence and pelvic organ prolapse leakage (down to the uterus or the

pelvic organs are the vagina) are very common in pelvic floor disorders and will be discussed in this study[5].

In the U.S., the prevalent rate for women having at least one pelvic disorder is one half. One ninth undergoes surgery for pelvic organ prolapse or urinary incontinence in her lifetime. Among those who undergo surgeries for pelvic organ prolapse or urinary incontinence in their lifetime, one third has another surgery for pelvic floor disorders within 5 years of the first[6]. Even though there may be other complication conditions for which the women were initially operated, the high rate of re-admission back to surgery may still relate to the recurrence risk of pelvic floor disorders.

The rate of pelvic floor disorders is increasing dramatically and is expected to increase even more over the next 30 years within the aging population[7]. It is extremely important now to understand the risk factors and to prevent one fourth of women (prevented fraction[24]) from developing pelvic floor disorders, which would have save 90,000 women per year from suffering these disorders in the future.

Measurement error:

Measurement error is the variation between measurements of the same quantity on the same individual. Relationships between physical activity and pelvic floor disorders have become the focus of many analytic studies. Given often relatively limited variation in the assessment in measurement of predictor variables, the results of observational studies largely depend on the correct measurement of physical activity and other potential risk factors. However, errors in measurement in the predictor variables lead to a biased and inefficient estimate of the relationship of the exposure to disease[8].

The following formula and equations are from the book "Measurement Error in Nonlinear Models: A Modern Perspective. Second Edition" by R. J. Carroll et al[17].

There are two types of measurement error: classical measurement error and Berkson measurement error. In symbols, let X_i be the true variable and W_i be the variable measured with error. The classical measurement error model is stated as:

$$W_i = X_i + U_i$$

In the classical measurement error model, W_i is an unbiased measure of X_i, so that U_i must have mean zero ($E(U_i|X_i)=0$). The error structure of U_i could be homoscedastic (constant) or heteroscedastic.

Another type of measurement error model is Berkson measurement error model. The Berkson measurement error model differs from the classical measurement error by

$$X_i = W_i + U_i$$

Where $E(U_i|W_i)=0$ so that the true measurement dose have more variability than the estimates dose.

In the PhActs study, the assessment of physical activity is likely subject to measurement error. In this study, women complete the two same questionnaires separated by four months. The measurement error that occurs when assessing a person's physical activity involves bias related to the actual activities performed and the true duration of time they were performed in addition to random variation. There may be various reasons for the variability between the two reports, such as bad memories etc. The larger the correlation coefficient between the two responses, the more reliable the measurement, assuming their difference is not different from zero on average. In this study, I will consider a correlation coefficient larger than 0.9 to be acceptable, indicating random

error instead of measurement error. Measurement error often occurs as the "flattened-slope" pattern[8], which means the patients who report higher intensity of physical activities in the first response may tend to report less intensity in the second response, or vice versa.

The effect of measurement error becomes much more complicated if the model contains more than one predictor variable measured with error. Besides the assessments regarding activity measurement, external studies from previous literature show that body mass index and caffeine intake may be potentially measured with error[8,9].

Bootstrapping

Bootstrapping is a computer-intensive, general purpose simulation approach in statistical inference by resampling from an approximate distribution. The choice of the approximate distribution is by checking the empirical distribution of the existing data[21]. If the observed data is approximately independent and identically distributed, the bootstrapping will construct a number of resamples by random sampling with replacement from the original dataset[22].

Boostrapping is recommended whenever the distribution of a statistic is complicated or unknown because of its independence of distribution. It is also working well when the sample size is insufficient for direct statistical inference. However, the sample size needed for bootstrapping is hard to define since the number of bootstrap samples recommended in literature has increased as available computing power has increased[23]. Ideally, the same as most of the statistical analysis, as the number of bootstrapping sample is increasing, the procedure itself will reduce the effects of random sampling errors and gave much closer estimation.

Regression calibration and Simulation extrapolation (SIMEX)

Nonlinear measurement error models usually begin with an underlying nonlinear model such as logistic regression for the response Y in terms of predictors, including the perfect measured predictors and those measured with errors[17]. In order to distinguish two kinds of predictors, we add subscripts to the predictor X. X_Z (sometimes Z alone can be used to represent X_Z) will be used to represent the predictors which for all practical purposes, are measured without error; and X_U are the true values of these variables. However, for some reason, those true values cannot be measured or observed accurately for all study subjects. The critical point of the measurement

error model is that we can measure or observe a variable X_W, (sometime W alone can be used to represent X_W) which substitutes for the true and unobserved variable X_U. These two are related but differ by a measurement error. The parameters in the model relating response variable Y and (X_Z, X_U) cannot, of course, be estimated directly by fitting Y to (X_Z, X_U), since X_U is not observable. The goal of measurement error modeling is to obtain nearly unbiased estimates of these parameters indirectly by fitting a model for Y in terms of (X_Z, X_W).

To identify and select methods to adjust for measurement error in our study, regression calibration and simulation extrapolation methods will be performed and compared in this report. In principle, regression calibration is a technique that corrects biases in regression results in situations where exposure variables are measured with error. It is a statistical method for adjusting point and interval estimates of effect obtained from regression models commonly used in epidemiology for bias due to measurement error in assessing nutrients or other variables[8].

Regression calibration is appropriate in the following two situations. First, it is appropriate when a gold standard is available in a main study/internal or external validation study and a linear measurement error model with constant variance applies[10,11]. The main issue in the validation study is that this kind of study doesn't have data on the outcome variable for the primary

regression models but focuses on the comparison of the error-prone measurement and the gold standard. The regression parameter of interest cannot be estimated in the validation study[12]. The validation study with gold standard will not be specified in detail. Instead, the validation study with replicate measures will be further discussed.

The second situation in regression calibration is that when the main study has an internal validation study design with replicate measurements, and the value of the variance of measurement error can be assumed[10,11].

The regression calibration method is a straightforward approach in that we need only to fit the regression model with the X_W and X_Z instead of X_U. Replication is going to be used in a calibration function for estimating X_U. This first step of regression results in a calibration function for estimating X_U. The unobserved covariates are then replaced by their predicted values from the calibration model in a standard analysis. Finally, the standard errors are adjusted and calculated by bootstrap to account for the estimation of the unknown covariates[13].

The following mathematical inference and equations are exactly those found in the paper "The Regression Calibration Method for Fitting Generalized Linear Models with Additive Measurement Error" by James W. Hardin et al.

With k_i replicate measurement, the measurement error variance may be estimated by[13]

$$\hat{\Sigma}_{uu} = \frac{1}{\sum_{i=1}^{n}(k_i - 1)} \sum_{i=1}^{n}\sum_{j=1}^{ki} (W_{ij} - \overline{W}_{i.}) (W_{ij} - \overline{W}_{i.})^T$$

Alternatively, the $\overline{W}_{i.}$ means the average over the k_i replicate measurements. The user may specify this variance matrix if it is known or estimated. The same situation in the non-replica studies, if there is only one error-prone variable, the measurement error variance is assumed to be known or the reliability is assumed to be known. If more than one error-prone variable, the situation is much more complex that the covariance matrix must be calculated and specified before regression calibration[13,14].

We need information to substitute for $X_{U.}$ The linear approximant to X_U given ($X_{Z,}$ X_W) is (here we use W and Z instead of X_Z and X_W):

$$\hat{X}_u \approx \mu x_u + \begin{pmatrix} \Sigma_{xx} \\ \Sigma_{zz} \end{pmatrix}^T \begin{pmatrix} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \overline{W} - \mu_w \\ Z - \mu_z \end{pmatrix}$$

In order to perform the calculation to obtain the linear approximant, we make the estimated substitutions for true variables. First, we use $\mu_W = \mu_{XU}$; substituting the mean of the replicate values for the mean of the unknown true covariates[13,14]. In addition,

$$\hat{\Sigma}_{zz} = \frac{1}{(n-1)} \sum_{i=1}^{n} (Z_{ij} - \overline{Z}_{i.}) (Z_{ij} - \overline{Z}_{i.})^T$$

as the usual analysis of variance estimate. We use

$$\hat{\Sigma}_{xz} = \frac{1}{\nu}\Sigma_{i=1}^{n} k_i (\overline{W}_{i.} - \hat{\mu}_w)(Z_i - \overline{Z}_.)^T$$

$$\hat{\Sigma}_{xx} = \frac{1}{\nu}\{\Sigma_{i=1}^{n} k_i (\overline{W}_{i.} - \hat{\mu}_W)(\overline{W}_{i.} - \hat{\mu}_W)^T\} - \frac{n-1}{\nu}\hat{\Sigma}_{uu}$$

Where $\nu = \Sigma_i k_i - \Sigma_i k_i^2 / \Sigma_i k_{i.}$, the estimated variance matrix for the unknown X_U is seen in two components due to the variance of X_U and the measurement error variance[13].

Up to this point, we can derive the estimated values for the unknown X_U and perform a standard analysis replacing X_U by X_W.

Simulation extrapolation (SIMEX) is a simulation based method, sharing simplicity of application the regression calibration, but with larger computation intensity[15]. The idea underlying the method is that the effect of measurement error can be determined by simulation. The method can be simply separated into two steps. The first one is a re-sampling stage. In the re-sampling step, more datasets made up from the existing dataset with additional measurement errors are generated. For each data set, the naïve estimate of the parameter can be estimated and the trend of estimated parameters versus the variance of the extra error can be established. The correct estimators of the parameters are obtained in the second stage by extrapolating this trend back to the case of no measurement error[15, 16].

Statistical Analysis

A self-administered lifetime physical activity questionnaire (the LPAQ) was developed, tested and found to be reproducible in women ages 39 to 65 to assess current activity[20]. The LPAQ is self-administered to determine physical activity during leisure time, household time and outdoor time for the past year and historical back to the age of menarche, separated into 5 categories including menstruation to age 21, age 22 to age 34, age 35 to age 50, age 51 to age 65 as well as the past year. Occupational activity is addressed separately in the PhActs study by a separate, validated occupation questionnaire in a lifetime span.

For each activity reported on the LPAQ and Occupational Questionnaire, a MET value will be assigned from the Compendium of Physical Activities to show intensity of each activity. The Occupation Questionnaire doesn't assign the MET value for each occupation; instead, a categorical scale (1 to 4) is assigned for a group of activities. The levels of occupational MET are consistent with descriptions listed in the Compendium of Physical Activities. The MET score will then be multiplied by the average number of hours per week reported in order to calculate MET hours per week. Each activity will be summed up according to the different age spans to get the accumulative activity intensity over the lifetime.

In addition to the standard calculation of MET hours per week, the Strenuous Activity Index will be used to calculate the average of hours per week spent in strenuous physical activity. The Strenuous Activity Index provides alternative scoring of each listed activity in the LPAQ and occupational levels as a function of each activity's hypothesized contribution to the development of POP or SUI. Only the time spent in activities level 3 or 4 will be counted as strenuous activity in the occupational report. The strenuous activity for the LPAQ will be assigned in different ways according to the activity description.

1) Dataset statistical summary: description and covariate selection criteria:

The datasets have separate sub-datasets, including medical history, physical exam, LPAQ physical activities (leisure, household, outside) and occupation activity. The study aim is to build up a logistic regression model to test the relationship between the log odds of POP or SUI diseases and the physical activity strenuousness. In our study, the total activity measurement includes both the LPAQ activity and occupational activity.

Each activity is recorded twice separated by 2 to 4 weeks, ideally. The data of patients who completed both first and second response are collected and used to correct for errors in measurement in physical activity. The datasets need for their analysis included 31 observations

in the POP study and 29 observations in the SUI study (Table 1). The intraclass correlation coefficient, used to measure the two measurements, reproducibility for a lifetime span, are 0.66 in the LPAQ, 0.75 for the overall activity including the LPAQ and occupational activity, and 0.81 for the strenuous activity in the POP study (Table 6). The corresponding intraclass correlation coefficients in the SUI study are 0.81, 0.78 and 0.61 (Table 7), which suggests substantial measurement error in this study. The correlation plots for LPAQ, LPAQ plus occupation, and strenuous activities are shown in Figure 1 to Figure 3 in the POP study (the similar correlation plots for the SUI study). The different MET hr/wk between two responses is not necessarily increasing as the average number of activities increases as shown in Figure 4 to Figure 5. Plots in the SUI study were similar.

The potential risk factors in POP or SUI study are age, ethnicity/race, education, smoking, caffeine intake, childbirth (vaginal vs cesarean deliveries), hysterectomy, hormonal therapy, medical conditions, healthy condition and body mass index. By checking the frequency of the each potential risk factor, the variables can be identified that have less than 1/3 of cases. If included in the logistic regression model, the software would have to stop by computational error. Therefore those variables will be excluded from the present analysis. In addition, there are some other variables such as number of cesarean deliveries, number of vaginal deliveries, whether

hormone replacement pills or patches are being taken right now and whether the hormone replacement therapy was taken before, which have missing values in some cells and they may be dropped automatically by the software when running logistic regression. So these three variables will also be excluded.

Among the remaining potential risk factors, the predictor variables Z that are assumed to have been measured without appreciable error are age, body mass index, schooling, ever pregnant or not, ever experienced menopause or not, health condition, prescriptive medications and caffeine intake (an indicator variable)[17].

Method of repeated physical activity measurement (first paper then paper again, first paper then web, first web then paper, or first web then web again) is another important variable in the PhActs study. We found there was no statistically significant difference among these four groups (Table 2) in this small preliminary dataset. Therefore, 'reproducibility study' group is not included in the final model of the present analysis. Larger sample size may be required to detect any differences and incorporate this variable into the model.

2) Tables for Descriptive Statistics

	Cases	Controls	Total number of observation
POP	10	21	31
SUI	8	21	29
Total unique patients			39

Table 1, Case and control distribution in POP and SUI studies

	Paper-web	Web-paper	Paper-paper	Web-web	P-value among four groups
POP	1	1	23	6	0.22
SUI	1	2	24	2	0.68

Table 2, Descriptive statistics for the reproducibility

Potential risk factor	Number with factor absent	Number with factor present
cough	31	0
heart attack	31	0
angina	31	0
sleep apnea	31	0
major depression	30	1
diabetes	30	1
smoking /number of cigarettes per day	29	2
high blood pressure	28	3
hysterectomy	27	4
Cancer	26	5
ever pregnant or not	26	5
Arthritis	25	6
cesarean delivery	24	2

Table 3. Frequency table for the potential risk factors for both POP and SUI study

Variable	N	Mean	Std Dev	Minimum	Maximum
Total MET hrs/wk for the lifetime LPAQ activity in Response 1	31	405	207.6	100.3	997.5
Total MET hrs/wk for the lifetime LPAQ activity in Response 2	31	374.2	196.6	85.5	739
Total MET hrs/wk for the lifetime LPAQ activity and occupational activity in Response 1	31	490.8	207.5	122.3	1107
Total MET hrs/wk for the lifetime LPAQ activity and occupational activity in Response 2	31	455.4	197.4	163.3	800.9
Total MET hrs/wk for the lifetime strenuous LPAQ activity and occupational activity in Response 1	31	229	137.7	6.6	694.1
Total MET hrs/wk for the lifetime strenuous LPAQ activity and occupational activity in Response 2	31	200.7	123.7	32.3	545.4

Table 4. Statistical summary for the lifetime LPAQ, LPAQ plus occupation and strenuous activity MET hr/wk for the **POP** study

Variable	N	Mean	Std Dev	Minimum	Maximum
Total MET hrs/wk for the lifetime LPAQ activity in Response 1	29	350.1	157.3	100.3	721.2
Total MET hrs/wk for the lifetime LPAQ activity in Response 2	29	336.5	186.5	85.5	739.0
Total MET hrs/wk for the lifetime LPAQ activity and occupational activity in Response 1	29	442.1	164.2	122.3	835.2
Total MET hrs/wk for the lifetime LPAQ activity and occupational activity in Response 2	29	425.7	211.6	163.3	1034.0
Total MET hrs/wk for the lifetime strenuous LPAQ activity and occupational activity in Response 1	29	193.9	92.3	6.6	355.2
Total MET hrs/wk for the lifetime strenuous LPAQ activity and occupational activity in Response 2	29	178.5	112.7	32.3	504.6

Table 5. Statistical summary for the lifetime LPAQ, LPAQ plus occupation and strenuous activity MET hr/wk for the **SUI** study

	web	paper
web	0.92	
	n=6	n=1
paper		0.74
	n=1	n=23

	correlation coefficient	standard error	95% CI
web-web	0.92	0.196	0.54, 1
paper-paper	0.74	0.147	0.45, 1

Table 6. Correlations between the first response and the second response in the **POP** study for all activities lifetime

	web	paper
web	0.94	
	n=6	n=1
paper		0.71
	n=1	n=23

	correlation coefficient	standard error	95% CI
web-web	0.94	0.171	0.60, 1

Table 7. Correlations between the first response and the second response in the **SUI** study for the strenuous activities lifetime

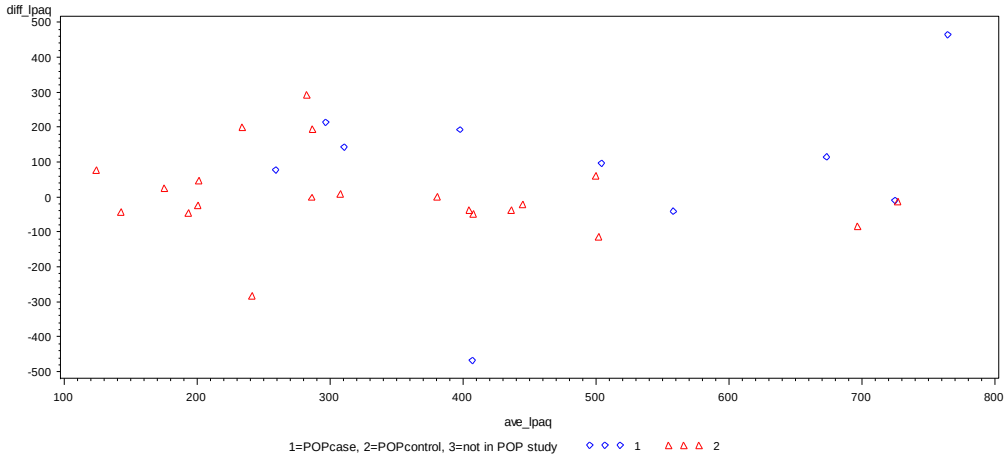

Figure 1. Bland and Altman plots of the difference in total MET hours lifetime between the two responses and the average MET hr/wk in lifetime LPAQ activities in the POP/SUI study ($p>0.05$ by t-test of the response 1 minus response 2 difference.)

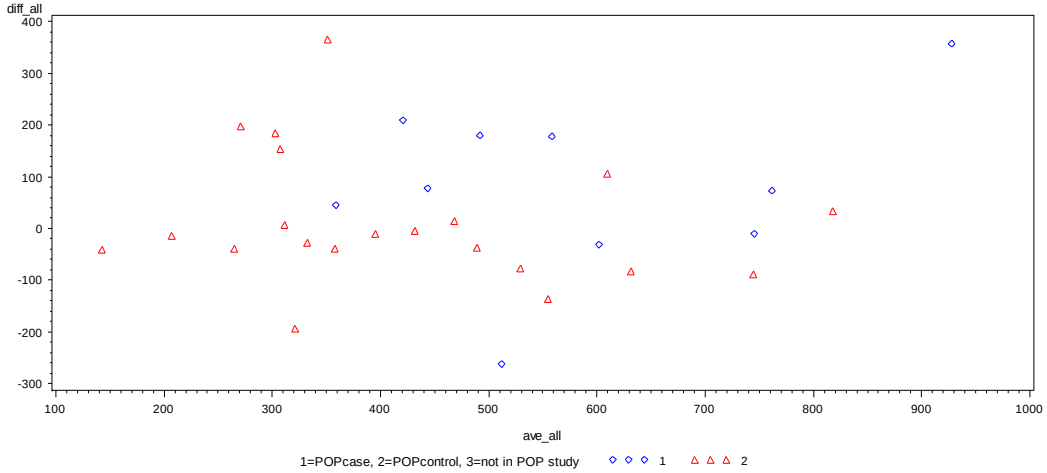

Figure 2. Bland and Altman plots of the difference in total MET hours lifetime between the two responses and the average MET hr/wk in the lifetime overall activities in the POP/SUI study. ($p>0.05$ by t-test of the response 1 minus response 2 difference.)

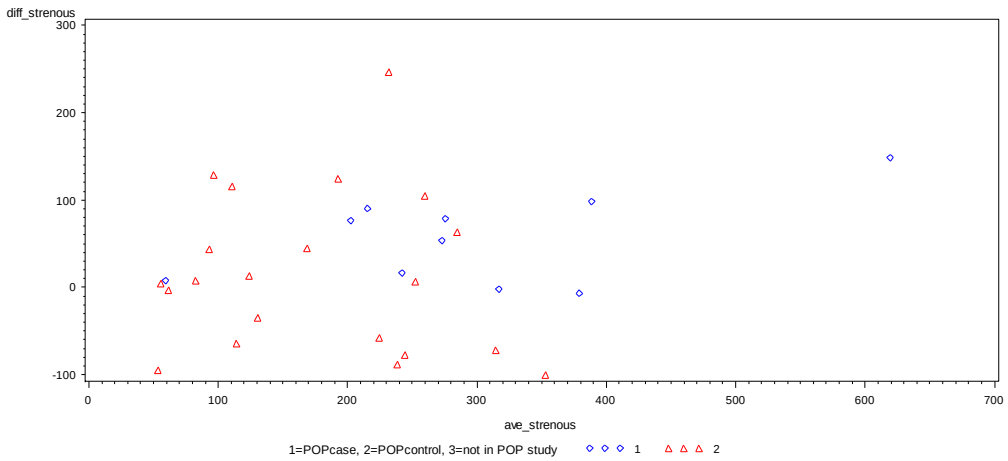

Figure 3. Bland and Altman plots of the difference in total MET hours lifetime between the two responses and the average MET hr/wk in the lifetime strenuous activities in the POP/SUI study

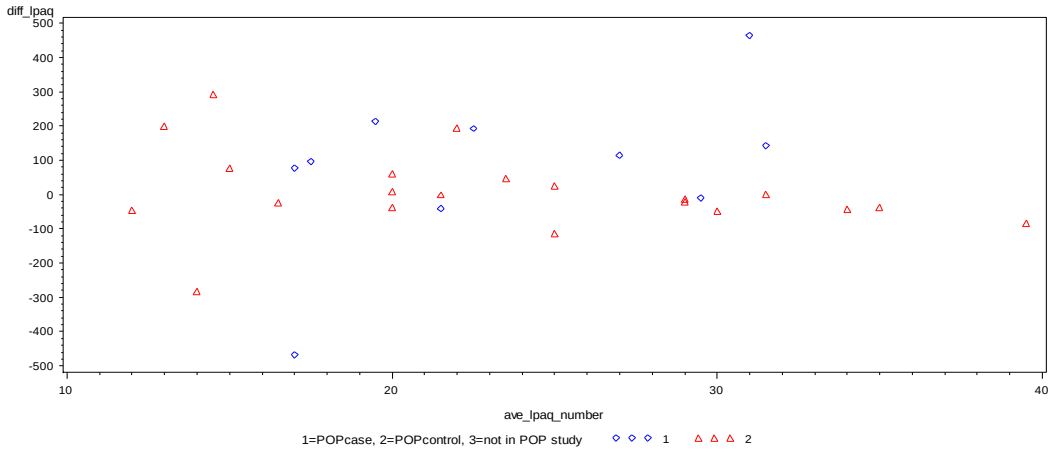

Figure 4. Bland and Altman plots of the difference in total MET hours lifetime between the two responses and the average numbers of activities in the lifetime LPAQ activities in the POP/SUI study

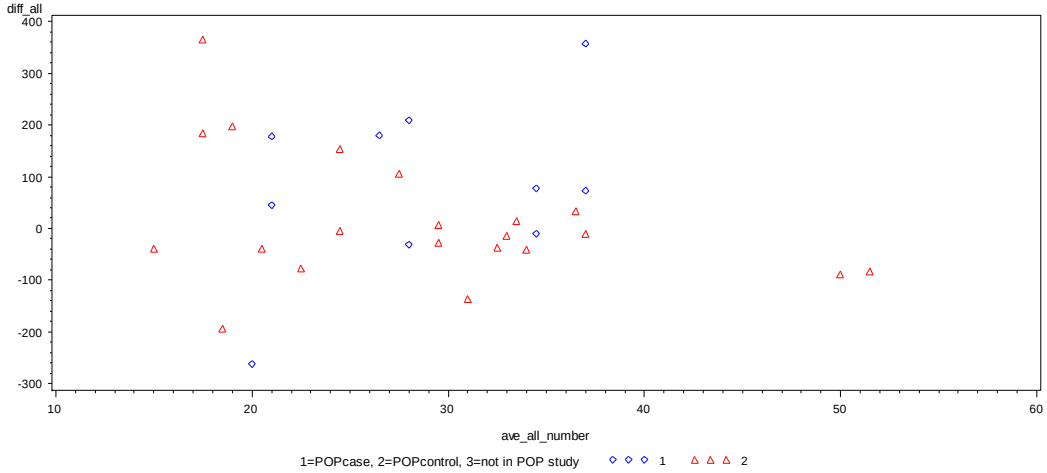

Figure 5. Bland and Altman plots of the difference in total MET hours lifetime between the two responses and the average numbers of all activities including LPAQ and occupation in the POP/SUI study

3) Regression calibration and Simulation extrapolation

In the subset of PhActs data used for the present analysis, repeat measurements of the physical activity MET hr/wk are obtained as response 1 and response 2. The internal repeat measurement approach is often more readily available in many large cohort studies[18]. However, not many case control studies such as ours are addressed by the literature. We will compare regression calibration with another method simulation extrapolation (SIMEX). Previous research indicated that physical activities were suggestive of a protective effect to some extent. However, too many or certain kinds of physical activities may lead to adverse effects[1,2].

Since there are so many variables compared to just a few observations, more variables are about to be dropped in the logistic regression model. Let's first take a look at the model with a binary outcome and each potential factor, one at a time. The activity measurement is the main focus and the study aim of the PhActs. So these are the variables which will be included for sure in the logistic regression model. By checking the correlation of activity measurement, we find that collinearity problem exists when both activity and strenuous activity are taken into account. Depending on the intraclass correlation coefficient, the correlation of coefficient is more likely to be really high. So each activity will be modeled separately. For a variable to have a qualitatively important effect, the odds ratio should be greater than 1.5 or less than 0.67. Under this standard,

the reasonable risk factors are menopause and health conditions chosen judging by their coefficients. Now we build a model with one continuous variable, activity; one binary variable, menopause; and one categorical variable, health condition as predictors. These same variables are included in the SUI study analysis.

By checking the linearity of outcome and predictors, we found that the log odd of POP/SUI was roughly linearly related to the activities intensity (Figure 6-Figure 9), which means the activity variable could be considered as the continuous predictor linearly related to the log odds of developing POP and SUI.

Here we compared the regular logistic regression model with the average activity as one covariate and menopause, health condition as the other two variables with the results from regression calibration and simulation extrapolation in the coefficient, standard error and 95% confident interval (Table 8 to Table 11). Although it is hard to make any conclusion according to the results shown below, as the number of observations is increasing, I am sure that the result will be much more convincing.

Using the SIMEX method, not only can we obtain the unbiased estimates and correct standard errors, we can also show that the amount of measurement errors affects the estimated coefficients.

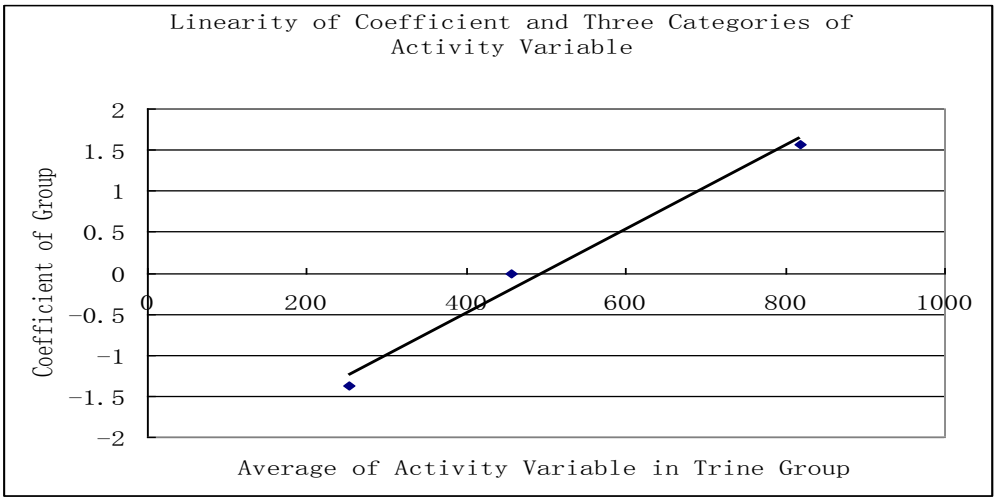

Figure 6. Linear relationship between the log odds of developing POP and lifetime overall activity at response 1.

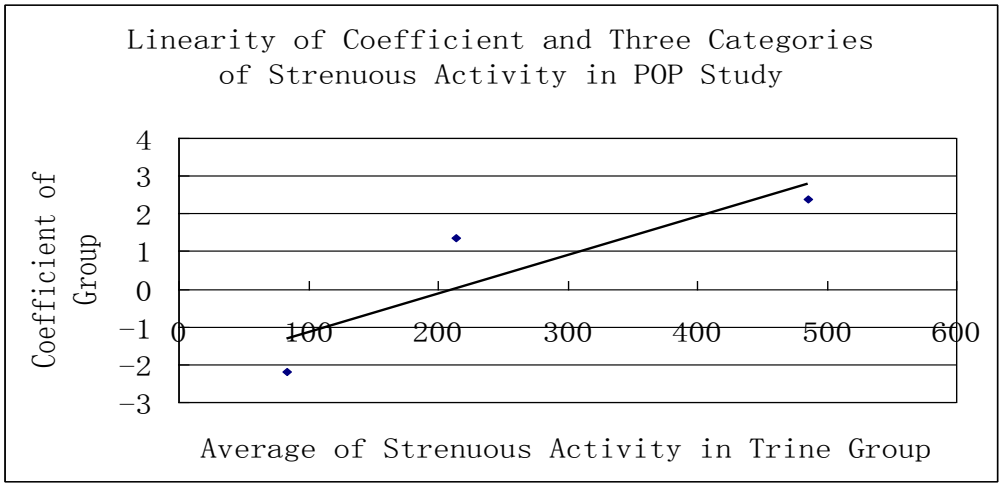

Figure 7. Linear relationship between the log odds of developing POP and lifetime overall strenuous activity at response 1.

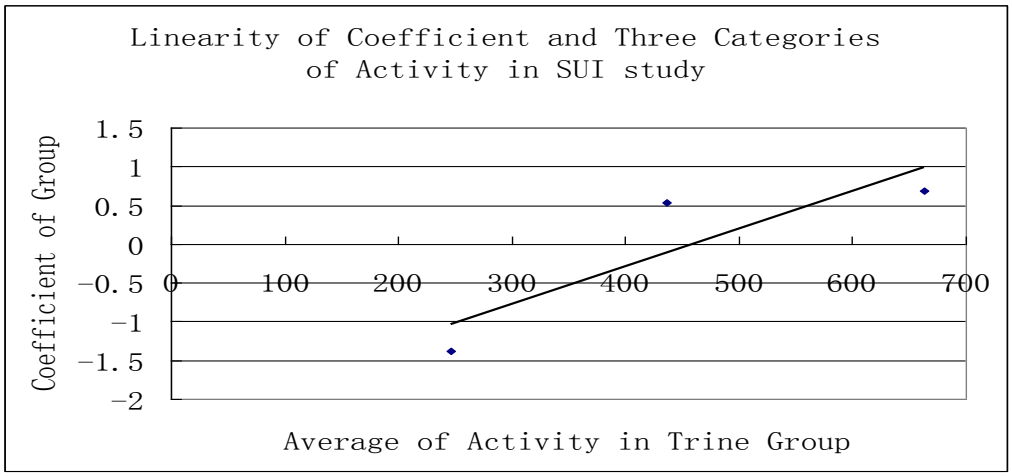

Figure 8. Linear relationship between the log odds of developing SUI and lifetime overall activity at response 1.

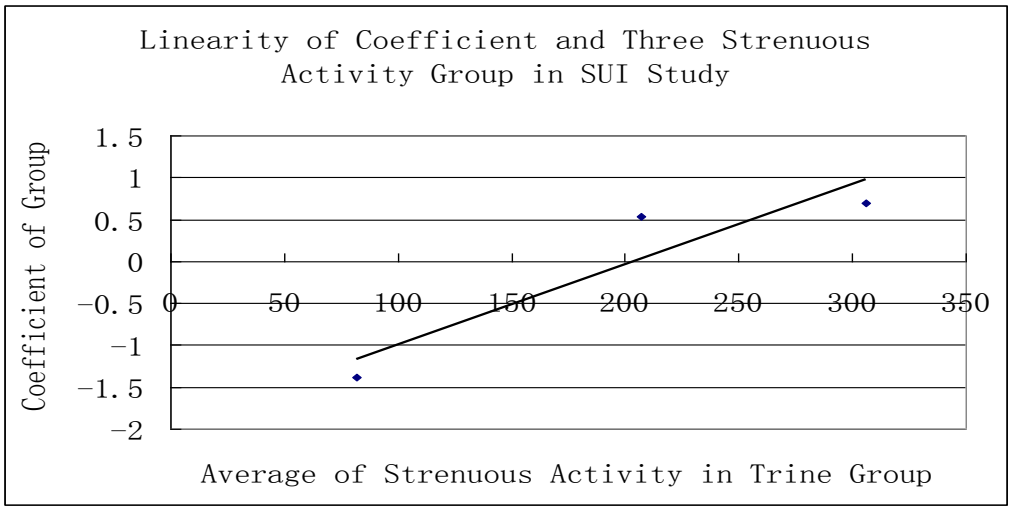

Figure 9. Linear relationship between the log odds of developing SUI and lifetime overall strenuous activity at response 1.

Replicate measurements variable	Coefficient	Standard Error	95% CI
Regular logistic regression	.0043	.00257	-.0007 .0093
Regression calibration	.0052	.00328	-.0016 .0119
Simulation extrapolation	.0044	.00981	-.0158 .0245

Table 8. Comparison of coefficient, standard error and 95% CI for measurement error correction in the lifetime activities among three methods in the POP study

Replicate measurements variable	Coefficient	Standard Error	95% CI
Naïve logistic regression	.0104	.00516	.0003 .0205
Regression calibration	.0119	.00627	-.0010 .0248
Simulation extrapolation	.0114	.02116	-.0322 .0549

Table 9. Comparison of coefficient, standard error and 95% CI for measurement error correction in the lifetime strenuous activities among three methods in the POP study

Replicate measurements variable	Coefficient	Standard Error	95% CI
Regular logistic regression	.0008	.00243	-.0040 .0055
Regression calibration	.0009	.00307	-.0054 .0073
Simulation extrapolation	.0008	.00650	-.0126 .0142

Table 10. Comparison of coefficient, standard error and 95% CI for measurement error correction in the lifetime activities among three methods in the SUI study

Replicate measurements variable	Coefficient	Standard Error	95% CI
Naïve logistic regression	.0047	.00494	-.0050 .0144
Regression calibration	.0065	.00736	-.0087 .0217
Simulation extrapolation	.0061	.01119	-.0170 .0290

Table 11. Comparison of coefficient, standard error and 95% CI for measurement error correction in the lifetime strenuous activities among three methods in the SUI study

The graph illustrates the extrapolated point estimates for activity fitted model in both lifetime activities and lifetime strenuous activities in **POP** study (Figure 7 and Figure 8)

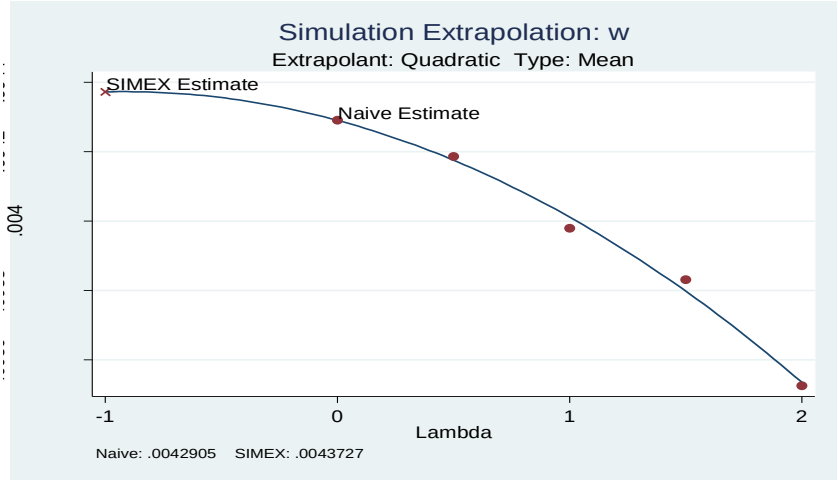

Figure 7. The effect of measurement error on parameter estimate in the lifetime activity

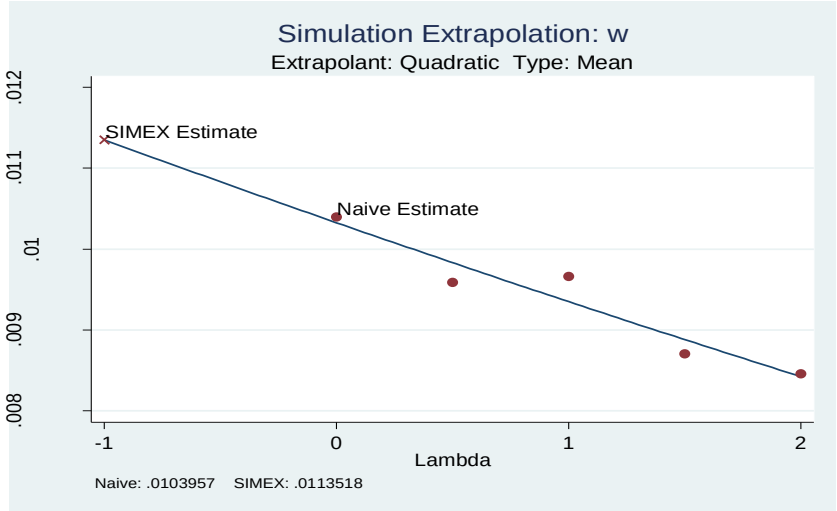

Figure 8. The effect of measurement error on parameter estimate in the lifetime strenuous activity

The graph illustrates the extrapolated point estimates for activity fitted model in both lifetime activities and lifetime strenuous activities in **SUI** study (Figure 9 and Figure 10)

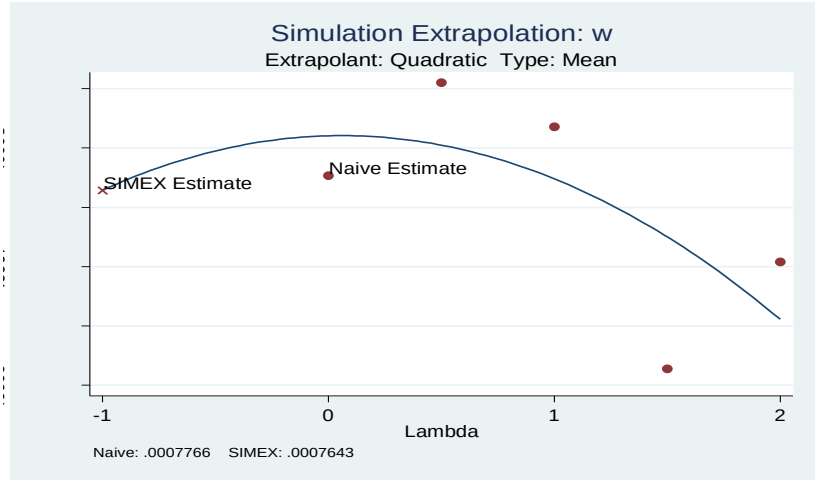

Figure 9. The effect of measurement error on parameter estimate in the lifetime activity

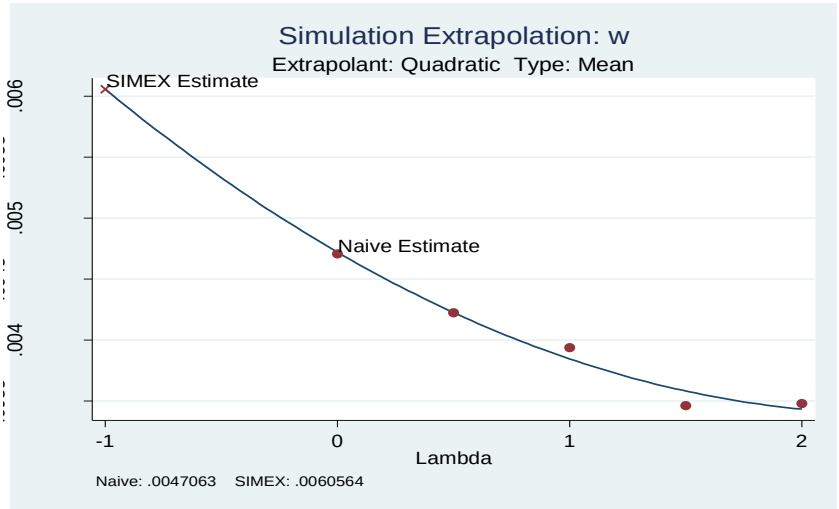

Figure 10. The effect of measurement error on parameter estimate in the lifetime strenuous activity

Discussion

The coefficient of average lifetime activity is 0.0043 in the regular logistic regression model, indicating the for 100 MET hours per week units increasing in lifetime, the expected change in the log of odds of developing POP disease is 100*0.0043=0.43. The odds ratio for developing POP for a patient with 100 more MET hours per week than another patient over their lifetime is exp(0.43)=1.53. After correction for error in measurement, The odds ratio for developing POP for a patient with 100 more MET hours per week than another patient over their lifetime is exp(0.52)=1.67 using regression calibration and exp(0.44)=1.55 using SIMEX.

The odds ratio for developing POP for a patient with 50 more strenuous MET hours per week than another patient over their lifetime is exp(50*0.0104)=1.68 using regular logistic regression model; exp(50*0.0119)=1.81 using the regression calibration; and exp(50*0.0114)=1.76 using SIMEX.

In this study, we consider the classical additive error model W=X+U where conditional on (Z, X) errors have mean zero and constant known covariance matrix[17]. The critical assumption for valid use of regression calibration in this study is that measurement error is non-differential with respect to the response variable, Y, that is, f(Y|x,X)=f(Y|X). In our study, the errors in

measurement are assumed to be independent from being chosen as a case or control in both POP and SUI studies. In the regression calibration study performed, the measurement error variance can be estimated using the formula in the introduction section (3662 for the strenuous activity and 10420 for all activity in the POP study; 4248 for the strenuous activity and 8578 for all activity in SUI study). Regression calibration is ideally suited for problems in which the calibration function E(X|W) can be estimated nearly unbiasedly and in generalized linear models such as logistic regression model in our case. Simulation extrapolation (SIMEX) is ideally suited for any problem with additive measurement error, and more commonly to those problems which can use the Monte Carlo methods to simulate the process of measurement error generation[17]. SIMEX is very general in the sense that the bias due to measurement error in almost any estimator of almost any parameter is readily estimated and corrected, at least approximately. The idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation. SIMEX assumes replicated measurements are available for all subjects, but no additional assumptions are made about the error variances, that is, it is not assumed that they are known and they could be homoscedastic or heteroscedastic[17]. In the present subset of PhActs data, all the observations have the record of completion of both response and measurements for the replica variable are available for all individual. However, the

full PhActs study was not planned to have replicates for every women. This means that SIMEX would be the method of choice, if the reliability of the activity variables from the reproducibility study is precisely measured. In conclusion, the regression calibration and simulation extrapolation both are suitable in addressing the errors in measurement in the PhActs study, but SIMEX may be financially more feasible. In addition, the results shown above need to be validated using larger sample size.

The bootstrapping method should give a better estimation of the parameter as well as the smaller standard error and more accurate 95% CI as the numbers of replication increase. However, it is not the case in this study (Table 12). The standard error in this study wasn't getting smaller as the number of bootstrapping replications increased, instead, the standard error showed as a fluctuating pattern. The main concern most likely to be related to this study is small sample size. Since the small sample size gives only a few values for the bootstrapping to choose from, so the bootstrapping sample will underrepresent the true variability since observations are frequently repeated and bootstrap samples, themselves, can repeat[25].

Number of Bootstrapping Replications	Coefficient of Activity	Standard Error	95% CI	
199 (default)	0.0046	0.00738	-.0105	.0198
1000	0.0047	0.00753	-.0108	.0202
2000	0.0052	0.00906	-.0134	.0238
5000	0.0051	0.00773	-.0108	.0210
10000	0.0045	0.00909	-.0142	.0232

Table 12. Standard error and 95% CI for different numbers of bootstrapping replications

The strategies I suggest to continue the PhActs study are first, much larger sample size is needed to have enough statistical analytical power (Table 13). The present analysis was limited by the need to exclude the variables due to small sample size. Especially when all the potential risk factors are needed to be considered in predicting the possibility of developing the POP or SUI disease, more observations are need to incorporate up to 15 covariates. Second, if it is financially impossible to collect the second response from every patient, external validation studies can be sought to obtain the variance of the measurement error required by regression calibration. Depending on what we have right now, in order to perform the analysis using regression calibration, another thought is to apply the measurement error we calculated to the whole dataset in the future, which overcomes the financially problem to collect every replicate measurement for all study population. Last, to improve the validity of statistical analysis, standardized data entry coding and data validation are not only necessary but important in the correct conclusion.

Intraclass correlation coefficient	Sample Size
0.75	674
0.80	147
0.85	50
0.90	21
0.95	8

Table 13. Correlation between two responses and sample size needed to make sure the 95% CI is at least above 0.7

Compared to the previous literature[20], our data shows the similar reliability between two measurements but in such a small sample size. Previous epidemiology studies have shown that physical activity is associated with a reduced incidence of disease. However, we are not able to draw this conclusion using current dataset. The result of this study shows that the average lifetime activity can be measured using a self-administered questionnaire. The reliability of the two replicated measurement is really high in our study.

Study limitation

In summary, statistical power may be adversely affected in the present analysis by a number of mechanisms:

1) A small study size, as expressed by the number of events, *Y*. There are only 31 observations with 10 cases and 21 controls in the pop study and even less (29 observations with 8 cases and 21 controls) in the SUI study. Further, the regression calibration and SIMEX won't take the observations with missing values into account, leading to only 16 ultimate usable observations finally. We addressed this limitation by reducing the number of potential risk factors in the logistic regression model to keep the numbers of observations still 31 and 29.

2) Multicollinearity between the predictor variables in the disease regression—the typical collinearity problem. For example, even though the variables whether hormone therapy was taken before and whether hormone therapy is taken right now are not in the final analysis because of missing values, however, if they were included, one of the two variables would be dropped due to the collinearity due to the perfectly correlated in the present subset of data. Also, the overall activity and strenuous activity assessment have high correlation with each other, leading to the collinearity problem.

3) Too many covariates compared to too few cases. Less statistical significance can be detected in this situation, and zero cells caused more variables to be dropped.

Designing a study of sufficient sample size to correct for the measurement error will help to diminish these types of problems by decreasing the variance of the calibration coefficients. If the validity correlation among replicates is low, the expected values, conditional on the replicated values and other variables measured without error, will often differ substantially from the actual individual true values. Finally, the regression calibration and simulation extrapolation analysis are commonly used and compared in the cohort study with large cohort numbers. More experience is needed to compare these methods in a case-control study.

Reference

1, Physical Activity and Weight Control. U.S. Department of Health and Human Services, National Institutes of Health

http://www.win.niddk.nih.gov/publications/physical.htm

2, Nygaard I, Handa V, et al. Changes in Physical Activity after Abdominal Sacrocolpopexy for Advanced Pelvic Organ Prolapse. *Am J Obstet Gynecol*. 2008 May ; 198(5): 570.e1–570.e5

3, Lianne F. Herbruck. The Impact of Childbirth on the Pelvic Floor: Vaginal Delivery

4, Rockwood TH, Church JM, Fleshman JW, Kane RL, Mavrantonis C, Thorson AG, Wexner SD, Bliss D, Lowry AC (1999). "Patient and surgeon ranking of the severity of symptoms associated with fecal incontinence: the fecal incontinence severity index." *Dis. Colon Rectum* 42 (12): 1525–32.

5, Pelvic Organ Prolapse

http://www.kmcnetwork.org/pelviccontrolcenter/organ_prolapse.cfm

6, Husam Abed, et al. Urinary Incontinence and Pelvic Organ Prolapse: Diagnosis and Treatment for the Primary Care Physician. *Med Clin N Am* 92 (2008) 1273–1293

7, K. Luber, et al. The demographics of pelvic floor disorders: Current observations and future projections. *American Journal of Obstetrics and Gynecology*, Volume 184, Issue 7, Pages 1496-1503

8, Anne C.M. Thiebaut, et al. Is It Necessary to Correct for Measurement Error in Nutritional Epidemiology? *Annals of Internal Medicine.* 2 January 2007 | volumn 146 Issue 1 | Pages 65-67

9, Joline WJ. Beulens et al. Alcohol Consumption and Risk for Coronary Heart Disease among Men with Hypertension. *Ann Intern Med*. 2007; 146:10-19

10, Donna Spiegelman et al. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statist. Med.* 2001; 20: 139-160

11, Donna Spiegelman et al. Regression calibration method for correcting measurement error bias in nutritional epidemiology. *Am J C/in Nutr* l997; 65(suppl): I 179S-86S

12, Roger Logan and Donna Spiegelman. The SAS %BLINPLUS Macro

13, James W. Hardin et al. The Regression Calibration Method for Fitting Generalized Linear Models with Additive Measurement Error. *The Stata Journal* (2003) 3, Number 4, pp. 361–372

14, James W. Hardin et al. Measurement error, GLMs, and notational conventions. *The Stata Journal* (2003) 3, Number 4, pp. 329–341

15, A. Guolo et al. A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. *Statistics in Medicine.* Volume 27 Issue 19, Pages 3755 – 3775

16, James W. Hardin et al. The simulation extrapolation method for fitting generalized linear models with additive measurement error. *The Stata Journal* (2003) 3, Number 4, pp. 373-385

17, R. J. Carroll et al. Measurement Error in Nonlinear Models: A Modern Perspective. Second Edition

18, May A Beydoun1, et al. Measurement error adjustment in essential fatty acid intake from a food frequency questionnaire: alternative approaches and methods. *BMC Medical Research Methodology* 2007, 7:41 doi:10.1186/1471-2288-7-41

19, Gary E. Fraser et al. Regression Calibration in Studies with Correlated Variables Measuredwith Error. *American Journal of Epidemiology* Vol. 154, No. 9 Pages 836-844

20, Lisa Chasan-Trber et al. Validity and reproducibility of a physical activity questionnaire in women. *Med Sci Sports Exerc.* 2002 Jun;34(6):987-92

21, Chernick, Michael R. (1999). Bootstrap Methods, A practitioner's guide. Wiley Series in Probability and Statistics.

22, Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC

23, Ying Zhang and Stephan Vogel, Significance tests of automatic machine translation evaluation metrics, *Machine Translation*, 10.1007/s10590-010-9073-6

24, John M. Last, A Dictionary of Epidemiology

25, Michael R. Chernick, Bootstrap Methods