Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sample Size Calculation for Prevalence Studies Using Scalex and ScalaR Calculators

Lin NAING (ayub.sadiq@ubd.edu.bn)
Universiti Brunei Darussalam Pengiran Anak Puteri Rashidah Sa'adatul Bolkiah Institute of Health Sciences https://orcid.org/0000-0003-1723-9854

Rusli NORDIN
Taylor's College - Subang Jaya Campus: Taylor's College Sdn Bhd

Hanif ABDUL RAHMAN
Universiti Brunei Darussalam Institute of Medicine: Universiti Brunei Darussalam Pengiran Anak Puteri Rashidah Sa'adatul Bolkiah Institute of Health Sciences

Yuwadi Thein NAING
Asia Pacific University of Technology & Innovation

Research Article

Keywords: Sample Size, Calculator, Single Proportion, Prevalence Studies

Posted Date: February 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1334522/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Sample Size Calculation for Prevalence Studies Using Scalex and ScalaR Calculators

Lin NAING*
Professor (Biostatistics), PAPRSB Institute of Health Sciences, Universiti Brunei Darussalam, Gadong, Brunei Darussalam
ORCID: 0000-0003-1723-9854

Rusli NORDIN
Professor (Epidemiology and Occupational Health) and Head of School, School of Medicine, Faculty of Health and Medical Sciences, Taylor's University, Subang Jaya, Malaysia
ORCID: 0000-0003-1878-3501

Hanif ABDUL RAHMAN
Lecturer (Biostatistics | Nursing), PAPRSB Institute of Health Sciences, Universiti Brunei Darussalam. Research Associate, Centre of Advanced Research (CARe), Universiti Brunei Darussalam, Gadong, Brunei Darussalam
ORCID: 0000-0003-3022-8690

Yuwadi Thein NAING
Graduate student, Asia Pacific University of Technology and Innovation, Malaysia
ORCID: 0000-0001-7842-0927

*** Corresponding author**
Lin NAING
Professor (Biostatistics),
PAPRSB Institute of Health Sciences,
Universiti Brunei Darussalam,
Jalan Tungku Link, Gadong,
Brunei-Muara BE3119
Brunei Darussalam
Email: ayub.sadiq@ubd.edu.bn

Key Words: Sample Size; Calculator; Single Proportion; Prevalence Studies

Word Count: 2192 (including abstract and references)

Declaration:

Ethics approval and consent to participate
The study did not require ethics approval and consent to participate.

Consent for publication
All authors have given consent for publication.

Competing interests
We do not have any competing interest.

Funding
This study is not funded by any funding agency.

Authors' contributions
LN, YTN contributed for designing and development of calculator.
All authors contributed for writing the paper.

Acknowledgements
No acknowledgment required.

Data and material availability
This paper doesn’t involve data. However, the free calculator is available here: (https://sites.google.com/view/sr-ln/ssc)

Abstract

Although books and articles guiding the methods of sample size calculation for prevalence studies are available, authors observe several issues related to sample size calculation in published articles such as incorrect calculation, incorrect use of formula, incorrect parameters, and incomplete sample size reporting. This brief article focuses on sample size calculation for a prevalence study, choosing correct parameters with proper understanding, and reporting issues, and demonstrate use of a well-prepared calculator which also assist users to make proper decision making and appropriate report. Two calculators can be used with free software (spreadsheet and RStudio) which benefit researchers with limited resources. The calculators are available at: (https://sites.google.com/view/sr-ln/ssc)

1. Introduction

In quantitative research, when we take a sample from a study population or eligible population in order to save our resources, there are two important statistical processes namely using a probability sampling method (commonly known as "random sampling") (Cochran, 1977), and calculating an appropriate sample size (Daniel & Cross, 2013). Both are equally important to ensure a good representative sample for the study population.

As we need a specific statistical analysis for a specific research objective, we also need a specific sample size calculation method for a specific research objective. Each objective requires a different sample size. In this paper, we focus on the objective that estimates a prevalence or proportion. For example, to estimate the prevalence of obesity, the prevalence of smoking, the prevalence of heart disease, diabetes mellitus or any other diseases of a study population. The method in this paper will not be suitable for other type of objectives such as estimating mean, comparing means, comparing proportions or regression analyses.

Although books and published articles guiding the methods of sample size calculation for prevalence studies are available, authors observe several issues related to sample size calculation in published articles such as incorrect calculation, incorrect use of formula, incorrect parameters, and incomplete sample size reporting. The worst is no sample size calculation at all.

Therefore, in this paper, authors address the correct method, correctly choosing parameters with a good understanding, adequate report for the publication, and use of a simple Excel calculator that guides users step-by-step and gives a draft report for publication. We believe that, this will improve sample size calculation in future prevalence studies in medical and health sciences.

2. Method to Calculate Sample Size

For an objective which estimates a prevalence, the sample size calculation formula is fairly simple and available in a number of books.

The following formula (Daniel & Cross, 2013) shall be used:

$$n = \frac{Z^2 P(1-P)}{d^2}$$

where n = Sample size,
Z = Z statistic for a level of confidence (1.96 for 95% confidence level),
P = Expected prevalence or proportion, and
d = Precision

However, we do not encourage researchers to use formula as it could have human error in manual calculation. We can use available software, and concentrate on carefully choosing appropriate parameters for the calculation.

2.1 Appropriately Choosing Parameters

The above formula indicates three parameters to be decided.

2.1.1 Parameter 1: Level of Confidence

When we take a sample but wish to know about the population (such as prevalence of smoking) from where the sample is taken, we will not know the exact prevalence of the population as we do not study all members of the population. However, the sample study gives us an estimation which has lower and upper limits (informally 'a range', but we call 'interval' in Statistics) for the population prevalence. We normally calculate these lower and upper limits or an interval with a certain level of confidence. Commonly used or almost always used "level of confidence" for these intervals or estimates, is 95% (which we called 95% confidence interval, CI) in medical and health fields. As this level of confidence is a very well-established common practice in medical and health fields, it is a straight-forward matter for our decision. Some software has fixed as 95% for the level of confidence without giving users' choice.

2.1.2 Parameter 2: Precision

As mentioned above, we will not know exact prevalence of the population as we do not study all members of the population. Therefore, the prevalence we calculate from the sample could deviate (error) from the population prevalence. We call this deviation as sampling error. We also know that, the larger the sample size, the smaller the errors in estimation will be. The errors are calculated as precision or also known as 'margin of error'.

Practically, the precision reflects the width of 95% confidence interval. If we decide to choose an absolute precision of ±2% in estimating a prevalence, we should expect, in the result, the width of 95% CI as 4% (e.g. 95% CI: 23%, 27%). If the absolute precision is ±5% in estimating a prevalence, we should expect, in the result, the width of 95% CI as 10% (e.g. 95% CI: 20%, 30%). The width of CI is twice that of the precision. Details are presented in Table 1.

Table 1: Relationship between Precision and width of Confidence Interval (CI)

Prevalence in Sample	Absolute Precision	95% CI for Population	CI Width	Required Sample size
25%	± 2%	(23%, 27%)	4%	1801
25%	± 5%	(20%, 30%)	10%	289
25%	±10%	(15%, 35%)	20%	73
30%	± 2%	(28%, 32%)	4%	2017
30%	± 5%	(25%, 35%)	10%	323
30%	±10%	(20%, 40%)	20%	81

It is an opportunity for researchers to decide precision (margin of error) and the width of CI that they wish to see in the results. Normally, researchers wish to have narrower width of CI but the narrower it is, the more expensive (bigger sample size) it is going to be. Even if researchers

decide to go for a smaller sample size, the researchers can also foresee or appreciate how poor CI width is going to be in their results. Therefore, this is an informed decision to be made by researchers.

2.1.3 Parameter 3: Variability of the Data

The larger variation the data has, the larger the sample size needed. This relationship can be explained in a simple analogy. When we cook soup and near to the finish, we stir it well before we taste. We always need a very small amount (small sample size) to taste because we stir it well and the variation is almost zero.

Practically in estimating the prevalence, the prevalence has effect on this variation and therefore effect on the required sample size. The relationship of prevalence and the sample size is presented in Figure 1.

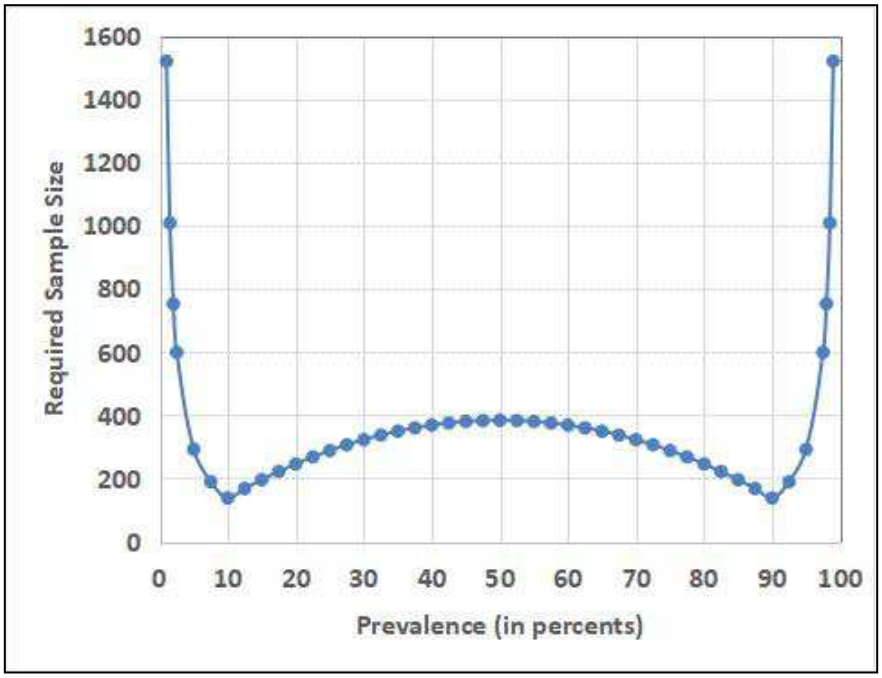


Figure 1: Prevalence and Effect on Sample Size

Obviously, it is the research objective to estimate the prevalence and researchers do not know this prevalence. Therefore, to calculate sample size, we normally find it out from most recent published studies with similar study population. If we cannot find suitable studies in the literature, we may consider to conduct a pilot study.

When we find multiple suitable prevalence from the literature, for example ranging from 15% to 30%, we should use the prevalence giving the highest sample size (in this case, 30%) in accordance with Figure 1 that shows that 30% will require the largest sample size in that range of 15% to 30% prevalence. Similarly, if the prevalence ranges from 60% to 80% in the recent literature, we should use 60% as it requires the largest sample size in that range.

We would like to caution that some books or guidelines suggest to use expected prevalence 50% if you could not get the prevalence at all (Daniel and Cross, 2013; Lwanga and Lemeshow 1991; Maple Tech, 2019). We discourage this practice. We should note from the example in Figure 1 that the prevalence of 50% will produce the largest sample size within the range of 10% and 90% of the prevalence. The required sample size is much higher in the region below 10% and above 90%. Therefore, a short cut of prevalence 50% may be used only if we are confident that the prevalence is expected to be between 10% and 90%.

2.1.4 Parameter 4: Anticipated Loss

We always have loss in sample size during the research process due to several reasons, such as non-response, incomplete data, loss-to-follow up, etc. Researchers should estimate for the loss with their past experience, and inflate the sample size in calculation accordingly.

2.2 Sample Size Calculation Report

The report of sample size should be reproducible. It means that all parameters used must be reported. There are four parameters namely, level of confidence (mostly 95%), expected prevalence (mostly from literature or pilot study), the precision or margin of error of estimate (decision by researchers) and anticipated loss (experience of researchers) used in the calculation. We should also include the name of software or calculator with proper reference. Scalex SP calculator has incorporated the draft report for the user to copy and use it. It ensures all necessary parameters used to be included in the report.

3. Application of Scalex SP and ScalaR Calculator

3.1 Simple Three Steps for Scalex SP

Basically, the Scalex SP calculator (Scalex stands for 'Sample Size Calculator using Excel', and SP stands for 'Single Proportion') (available at: https://sites.google.com/view/sr-ln/ssc) guides the users in three steps:
Step 1: to type in "Expected Prevalence" in terms of per cent (>0 to <100)
Step 2: to type in "Anticipated Loss" in terms of per cent (0 to <100)
Step 3: to decide and type in the precision of user choice after going through the Sample Size Table. Users may type a precision which is not listed in the table (e.g. ±2.5%). Then, Scalex SP will give a draft report for the user.

Major advantage of the Scalex SP calculator is that, it gives users Sample Size Table (Figure 3) in which users can appreciate sample sizes for a range of precision, and appreciate or foresee the CIs in their results. Therefore, it helps users in decision making of selecting precision considering available resources.

3.2 Example using Scalex SP

We are going to conduct a study to estimate prevalence of obesity among secondary school children in a district. We managed to find the expected prevalence in the literature as 30%.

When we start the Scalex SP, we see the interface as in Figure 2. Then, we fill 30 (30%) for Expected Prevalence. As we experienced 10% non-response in this study population in previous studies, we fill 10% loss (see Figure 3).

Then, sample sizes given for various precisions are reviewed and we decide to use ±3% precision as it gives us an acceptable width of 95% CI (27%, 33%), and the sample size (*n*=997) is possible to manage.

Then, we fill in 3 (3%) in Step 3, and Scalex SP gives the draft report as in Figure 3.

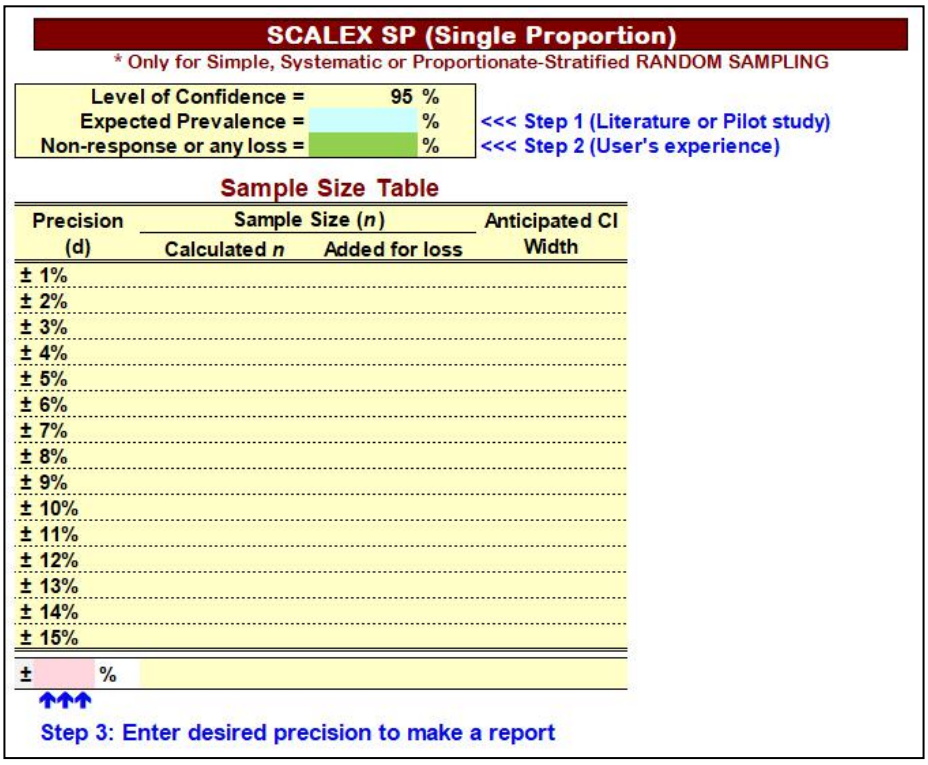
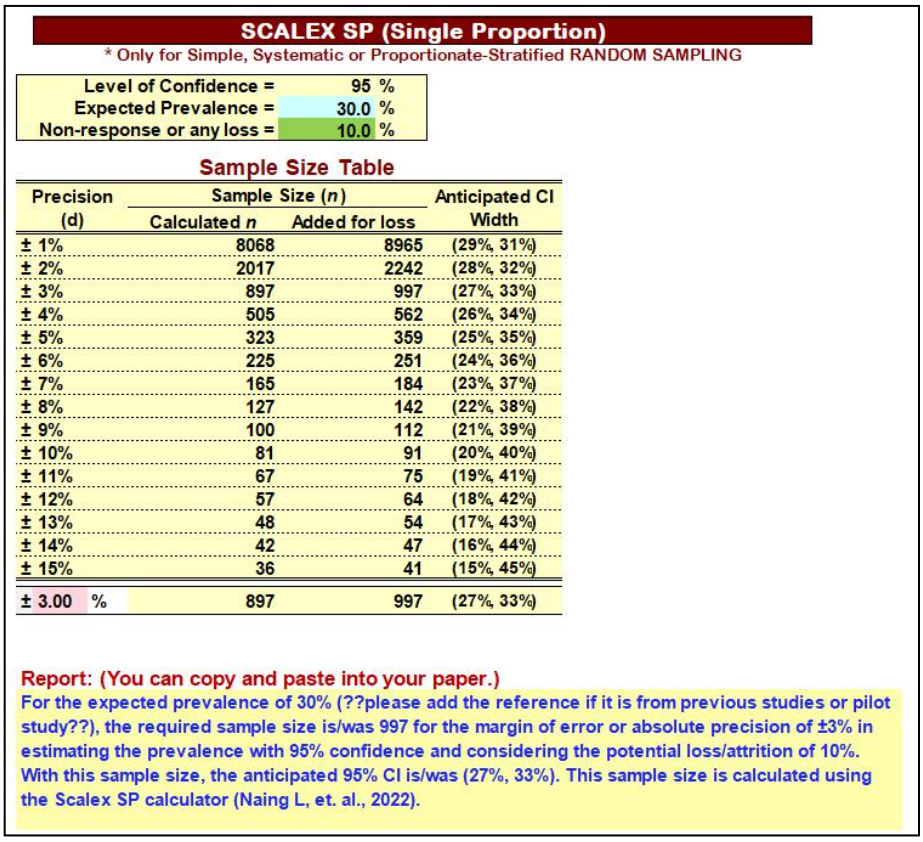


Figure 2: Scalex SP interface for Step 1, 2 and 3



SCALEX SP (Single Proportion)

* Only for Simple, Systematic or Proportionate-Stratified RANDOM SAMPLING

Level of Confidence =	95 %
Expected Prevalence =	30.0 %
Non-response or any loss =	10.0 %

Sample Size Table

Precision (d)	Sample Size (*n*) Calculated *n*	Added for loss	Anticipated CI Width
± 1%	8068	8965	(29%, 31%)
± 2%	2017	2242	(28%, 32%)
± 3%	897	997	(27%, 33%)
± 4%	505	562	(26%, 34%)
± 5%	323	359	(25%, 35%)
± 6%	225	251	(24%, 36%)
± 7%	165	184	(23%, 37%)
± 8%	127	142	(22%, 38%)
± 9%	100	112	(21%, 39%)
± 10%	81	91	(20%, 40%)
± 11%	67	75	(19%, 41%)
± 12%	57	64	(18%, 42%)
± 13%	48	54	(17%, 43%)
± 14%	42	47	(16%, 44%)
± 15%	36	41	(15%, 45%)
± 3.00 %	897	997	(27%, 33%)

Report: (You can copy and paste into your paper.)

For the expected prevalence of 30% (??please add the reference if it is from previous studies or pilot study??), the required sample size is/was 997 for the margin of error or absolute precision of ±3% in estimating the prevalence with 95% confidence and considering the potential loss/attrition of 10%. With this sample size, the anticipated 95% CI is/was (27%, 33%). This sample size is calculated using the Scalex SP calculator (Naing L, et. al., 2022).

Figure 3: Scalex SP with Report

3.3 ScalaR Programme for R users

Authors have written R Script (ScalaR SP.R) and with two command lines as in Figure 4 (this Script file must be stored at "Working Directory"), will give the same output as Scalex SP. (available at: https://sites.google.com/view/sr-ln/ssc)

Example of R command as:
> ScalarSP(p=0.3, d=0.03, loss=0.1)
p= expected prevalence
d = precision or margin of error
loss = anticipated loss or attrition of sample size

```
> source("ScalaR SP.R")
> scalarSP(p=0.3,d=0.03,loss=0.1)

Sample Size Report begins ....
============================================================

Table 1: Sample Size Table for Expected Prevalence 30%
------------------------------------------------
   Precision n    n++  95%CI
1  ±1%       8068 8965 (29%, 31%)
2  ±2%       2017 2242 (28%, 32%)
3  ±3%       897  997  (27%, 33%)
4  ±4%       505  562  (26%, 34%)
5  ±5%       323  359  (25%, 35%)
6  ±6%       225  251  (24%, 36%)
7  ±7%       165  184  (23%, 37%)
8  ±8%       127  142  (22%, 38%)
9  ±9%       100  112  (21%, 39%)
10 ±10%      81   91   (20%, 40%)
11 ±11%      67   75   (19%, 41%)
12 ±12%      57   64   (18%, 42%)
13 ±13%      48   54   (17%, 43%)
14 ±14%      42   47   (16%, 44%)
15 ±15%      36   41   (15%, 45%)
------------------------------------------------
Precision = Margin of Error; n++ = added for 10% loss
CI = Confidence Interval


------------------------------------------------
Draft Report (You can copy and paste in your report)
For the expected prevalence of 30% (please add the reference if it is
from previous study), the required sample size is/was 997 for the
margin of error or absolute precision of ±3% in estimating the
prevalence with 95% confidence and considering a potential
loss/attrition of 10%. With this sample size, the anticipated 95% CI
is/was (27%, 33%). This sample size is calculated using the ScalaR SP
(Naing L, et. al., 2022).
[End of report]
```

Figure 4: ScalaR SP - with report

3.4 Other Issues

The Scalex calculator is for the studies using the specific sampling methods such as simple random sampling, systematic sampling, and proportionate-stratified random sampling. For other sampling methods, the calculated sample size should be multiplied with the design effect (Lwanga and Lemeshow 1991). Estimating design effect could be from the literature if it is reported in the previous similar studies. If not, it is a complicated procedure involving data simulation.

With regard to potential loss or attrition, although we can put any per cent of the potential loss and inflate the sample size, it doesn't guarantee that the calculated sample size is valid in term of representative sample. There are different opinions on the acceptable per cent of loss or attrition

(Draugalis and Plaza, 2009) depending on the type of studies, it is important to note that the higher the loss or attrition, the larger the compromise of the validity of results.

4. Conclusion

With technological advancement, researchers should not calculate sample sizes manually. The software or calculators should help researchers minimize possible error in calculation and also to assist in reporting. However, the use of correct parameters still remains as the responsibility of users. In addition, calculators using free software, will benefit researchers who have limited resources. It is available at: (https://sites.google.com/view/sr-ln/ssc). Authors will continue to use Scalex calculator for other type of studies in the near future.

5. References

Cochran, W. G. (1977), *Sampling Techniques*, 3rd Edition, New York, John Wiley and Sons

Daniel, W. W., and Cross, C. L. (2013), *Biostatistics: A Foundation for Analysis in the Health Sciences,* 10th Edition, New York: John Wiley and Sons

Lwanga, S. K., and Lemeshow, S. (1991), *Sample Size Determination in Health Studies: A Practical Manual,* Geneva: World Health Organization

Maple Tech, I. L. (2019), Calculator.net [online], Retrieved December 19, 2019, Available at https://www.calculator.net/sample-size-calculator.html?type=1&cl=95&ci=5&pp=50&ps=&x=120&y=21.

Draugalis, J. R. and Plaza, C. M. (2009), "Best Practices for Survey Research Reports Revisited: Implications of Target Population, Probability Sampling, and Response Rate," American Journal of Pharmaceutical Education, 73(8), 142.

6. Author Contact Information

First & corresponding author:
Lin NAING
Professor (Biostatistics)
PAPRSB Institute of Health Sciences
Universiti Brunei Darussalam
Jalan Tungku Link, Gadong, BE 1410
Brunei Darussalam
Tel: +673 2463001
e-mail: ayub.sadiq@ubd.edu.bn; naing61@gmail.com

Other authors:

Rusli NORDIN
Professor (Epidemiology and Occupational Health) and Head of School, School of Medicine,
Faculty of Health and Medical Sciences,
Taylor's University,
Subang Jaya,
Malaysia
e-mail: Rusli.Nordin@taylors.edu.my

Hanif ABDUL RAHMAN
Lecturer (Biostatistics | Nursing), PAPRSB Institute of Health Sciences
Research Associate, Centre of Advanced Research (CARe)
Universiti Brunei Darussalam
Jalan Tungku Link, Gadong, BE 1410
Brunei Darussalam
e-mail: hanif.rahman@ubd.edu.bn

Yuwadi Thein NAING
Graduate student, Asia Pacific University of Technology and Innovation, Malaysia
e-mail: jayjay.naing@gmail.com

Acknowledgments and relevant information
No acknowledgment and other relevant information