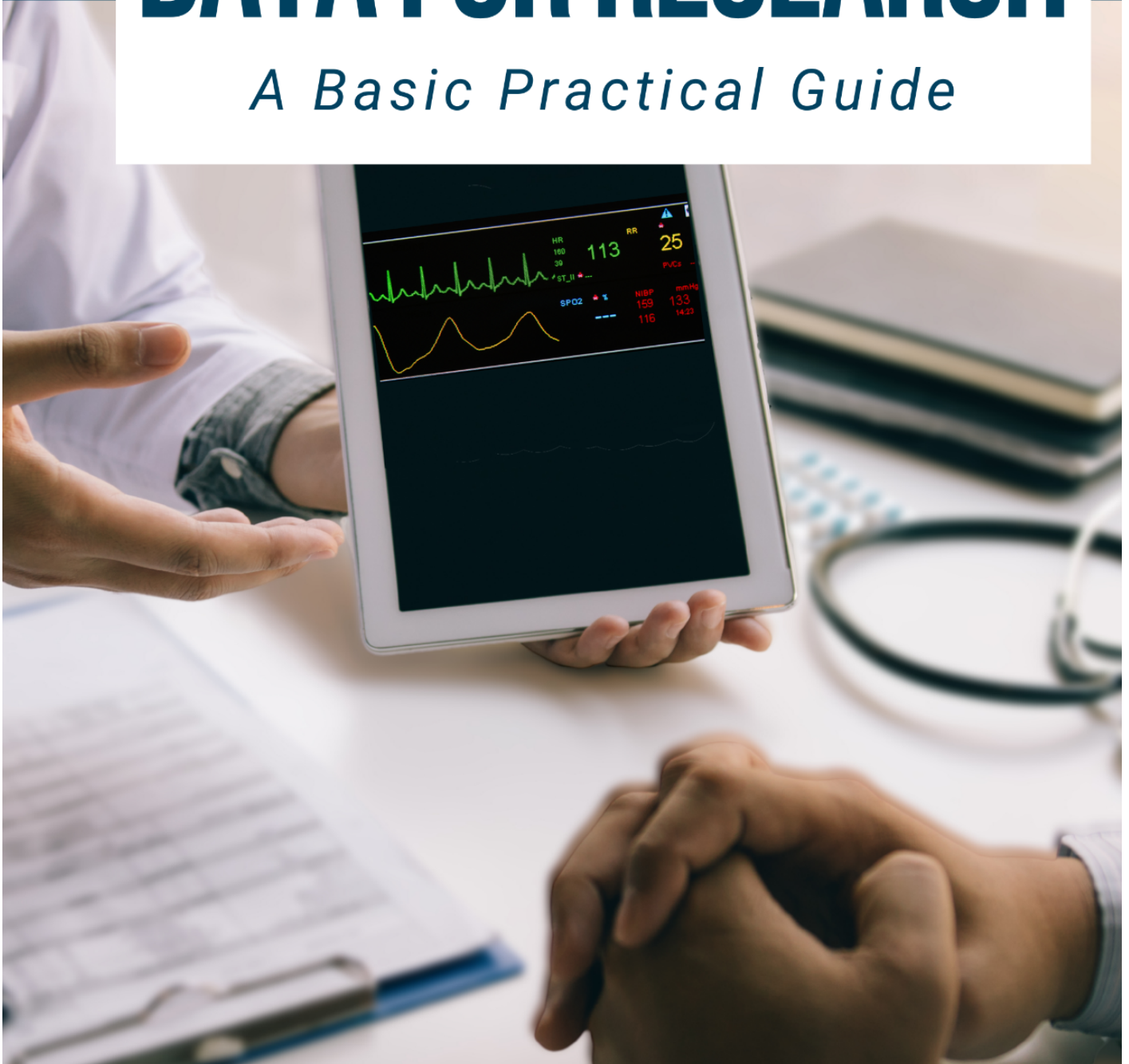


MOHAMAD ADAM BUJANG

# PATIENT REGISTRY DATA FOR RESEARCH

*A Basic Practical Guide*



EDITORS:

PROF. DR. JAMALLUDIN AB RAHMAN &  
PROF. DR. SHAMSUL AZHAR SHAH

**A BOOK BY INSTITUTE FOR CLINICAL RESEARCH, NIH**

**MOHAMAD ADAM BUJANG**

***Patient Registry Data for Research: A Basic  
Practical Guide***

*With*

**ANG SWEE HUNG, TG MOHD IKHWAN TG ABU BAKAR SIDIK,  
TASSHA HILDA ADNAN, NADIAH SA'AT, HON YOON KHEE &  
ALAN FONG YEAN YIP**

*Editors*

**PROF. DR. JAMALLUDIN AB RAHMAN**

**&**

**PROF. DR. SHAMSUL AZHAR SHAH**

**May 2021**

Institute for Clinical Research (ICR),  
National Institutes of Health, Blok B4, Aras 1, Jalan Setia Murni U13/52, Seksyen U13, Setia  
Alam, 40170 Shah Alam, Selangor, Malaysia  
[www.crc.gov.my](http://www.crc.gov.my) | Phone: 603-3362 7700

© INSTITUTE FOR CLINICAL RESEARCH (ICR), Mohamad Adam Bujang, 2021.

Cover and interior design by Chew Cheng Hoon.

All rights reserved.

The authors retain the sole copyright ownership of this book. Therefore, it is necessary to obtain prior permission from either the author or the Director of Institute for Clinical Research, Ministry of Health, Malaysia, in order to reuse the content(s) of this book and/or to reproduce the content(s) of this book in any form. However, the author wishes for this book to serve as a useful and not-for-sale reference for managing and conducting medical research consultations.

## **PATIENT REGISTRY DATA FOR RESEARCH: A BASIC PRACTICAL GUIDE**

First Edition, 2021.

Cite as: Mohamad Adam B., Swee Hung A., Tassha Hilda A., Tg Mohd Ikhwan T.A.B.S., Nadiah S., Yoon Khee H., Alan Fong Y.Y. (2021) PATIENT REGISTRY DATA FOR RESEARCH: A BASIC PRACTICAL GUIDE. Kuala Lumpur, Malaysia: Institute for Clinical Research, NIH MY.

DOI: [doi.org/10.5281/zenodo.4722674](https://doi.org/10.5281/zenodo.4722674)

ISBN: 9781005767600



## **Authors**

Dr. Mohamad Adam Bujang, PhD (Information Technology & Quantitative Sciences)  
*Clinical Research Centre, Sarawak General Hospital*  
*Institute for Clinical Research*  
*Ministry of Health Malaysia*

Dr. Ang Swee Hung, Msc (Public Health)  
*Institute for Clinical Research*  
*Ministry of Health Malaysia*

Mr. Tg. Mohd Ikhwan Tg Abu Bakar Sidik, Msc (Statistics)  
*UKM Medical Molecular Biology Institute*  
*Universiti Kebangsaan Malaysia*

Mrs. Tassha Hilda Adnan, Bsc (Statistics)  
Sector for Biostatistics & Data Repository,  
*National Institutes of Health*  
*Ministry of Health Malaysia*

Ms Nadiah Sa'at, Bsc (Mathematics)  
*Institute for Clinical Research*  
*Ministry of Health Malaysia*

Mr. Hon Yoon Khee, BPharm (Pharmacy)  
*Institute for Clinical Research*  
*Ministry of Health Malaysia*

Dr Alan Fong Yean Yip, MRCP (UK), FRCP (Edin)  
*Clinical Research Centre, Sarawak General Hospital*  
*Institute for Clinical Research*  
*Ministry of Health Malaysia*

## **Editors**

Prof. Dr. Jamalludin Ab Rahman  
*Deputy Dean (Postgraduate & Research)*  
*Department of Community Medicine*  
*Kulliyyah (Faculty) of Medicine*  
*International Islamic University Malaysia*

Prof. Dr Shamsul Azhar Shah  
*Community Health Department*  
*Faculty of Medicine,*  
*University Kebangsaan Malaysia*

## ***Acknowledgements***

*The authors would like to thank the Director-General of Health Malaysia for his permission to publish this e-book/book. The appreciation also goes to Dr. Lim Teck Onn, Dato' Dr. Goh Pik Pin, the former Directors of Institute for Clinical Research, and Dr Kalai Peariasamy, the Director of Institute for Clinical Research, for their support and guidance. A special appreciation also goes to Dr Chew Cheng Hoon for helping in the design of the book's cover and also in the process to publish this e-book/book. Last but not least, the author would like to express his sincere appreciation to our colleagues, Mr. Abd Muneer Abd Hamid, Mdm. Premaa Supramaniam, Mdm. Evi Diana Omar, Mr. Shahrul Aiman Soelar, Mdm. Nor Aizura Zulkifli, Mdm. Mariana Mohamad Ali, Mdm. Nur Amirah Zolkepali and Mdm. Nurakmal Baharum for their involvement in conducting the statistical analysis service for producing reports and scientific articles from patient registry databases for Institute for Clinical Research's research clients.*

# *Table of Content*

Preface

Abbreviations

Tables

Figures

Chapter 1 - Background

1.1 Purpose and Scope

Chapter 2 - Data Acquisition

2.1 Obtain Prior Consent and Study Approval

2.2 Understand the Scope of a Registry

2.3 Estimate the Minimum Sample Size Required for a Registry Study

2.4 Prepare a Dummy Table for Statistical Analysis

Chapter 3 - Data Preparation

3.1 Proper Handling of Duplicates Found in a Dataset

3.2 Match and Combine Data from Different Datasets

3.3 Set up Conditions and Requirements for Analysis

3.4 Establish Data Cleaning Procedures

3.5 Proper Handling of Missing Data in a Registry Database

3.6 A Summary of Data Preparation Procedures

Chapter 4 - Approach for Statistical Analysis

4.1 A Proposed Guide for Statistical Analysis of Patient Registry Data

Before statistical analysis

During statistical analysis

After statistical analysis

4.2 Presentation of the Final Results

4.3 Development of the Statistical Analysis Plan

Chapter 5 - Summary

References

Appendix 1

Examples list of scientific articles published based on data from patient registries by the authors

Appendix 2

## Example part of a Case Report Form (CRF) from National Cardiovascular Disease Database (NCVD)

About the authors



## *Preface*

Analysis of patient data can be a complicated and challenging process, especially when the data involve many subjects and many variables. A patient registry is a database that organizes collecting the important set of data on a list of identifiable individuals for a specific disease. This type of data usually has tons of data and hundreds of different variables. Thus, the approach to conducting research by using a patient registry database will be more complicated than the other types of dataset. Since the handling of patient registry data is a challenging task, the authors have come out with this e-book/book to become a guideline for the statisticians, medical officers and scientists for them to refer as a handbook whenever they need to use patient registry data for their research.

The objective of this e-book/book is to describe a basic practical guide on conducting research using the patient registry. It is a guideline that has been drawn up from the wealth of practical experience in conducting registry-based research and a widespread consensus of various statisticians and clinicians. This guideline emphasizes data acquisition, data preparation and approach for statistical analysis. It also includes a checklist that summarizes a list of pertinent points for consideration by a novice researcher before he/she plans to embark on a registry-based study. The checklist can be used as a tool to guide all researchers, especially statisticians and clinicians, to plan and conduct a research study by using data from a patient registry.

The ultimate aim of this e-book/book is to provide a standardization regarding the approach to conducting research by using patient registry data. The benefits of using the checklist provided by this e-book/book are to avoid problems that are commonly encountered by researchers when they are conducting a registry-based study, such as failure to identify pertinent ethical issues when handling patients' confidential data, inadvertently obtaining invalid study findings due to improper and/or inadequate data preparation for statistical

analysis and adopting the wrong approach to data analysis. As the scope of the data preparation and the statistical analysis of a patient registry can be overwhelmingly huge. Although this e-book/book does not provide complete coverage of the subject matter, it is our fervent hope that the e-book/book will be the first point of reference for statisticians, clinicians and scientists when they are conducting a registry-based study by using patient registry data.

*The Authors*

*March 2021*

## *Abbreviations*

ACS - Acute Coronary Syndrome

ADCM - Adult Diabetes Control and Management

BMI - Body Mass Index

CD - Compact Disc

CRF - Case Report Form

DiCARE - National Database on Children and Adolescent with Diabetes

FAQs – Frequently Asked Questions

IC - Identity Card

IT - Information Technology

MOH - Ministry of Health

MNAR - Missing Not at Random

MCAR - Missing Completely at Random

NCVD - National Cardiovascular Disease Database

NED - National Eye Database

NDR - National Diabetes Registry

NED - National Eye Database

NGO - Non-Government Organization

NORM - National Orthopaedic Registry Malaysia

NSRM - National Suicide Registry Malaysia

SAP - Statistical Analysis Plan

US - United States

WHO - World Health Organization

## Tables

Table 3.1: Summary table for FAQ on handling duplicates

Table 3.2: Summary table for FAQ on handling data cleaning

Table 3.3: Summary table for FAQ on handling missing data

Table 5.1: Checklist for data acquisition, data preparation and approach for statistical analysis for research using patients' registries databases

## Figures

Figure 1.1: A common step-by-step process for conducting research using secondary data and the focus of this e-book/book

Figure 2.1: An illustration of a template dummy table versus an actual dummy table filled up with study results

Figure 3.1: Visual example and definition for duplicates, missing values, inconsistency and extreme values

Figure 3.2: Sample of problematic data

Figure 3.3: A sample of a dataset in excel sheet based on Obstructive Sleep Apnea (OSA) study

Figure 3.4: Variable definition of a sample of a dataset in excel sheet based on Obstructive Sleep Apnea (OSA) study (as shown in Figure 3.3)

Figure 5.1: A practical guide consisting of a list of recommendations for planning a research study by using data from patient registries

## Chapter 1 – Background

Clinical researchers aim to make consistent efforts to continuously search for new and improved methods to fight against all human diseases in order to achieve better clinical outcomes for mankind. Therefore, it is mandatory for researchers in the medical field to continuously improve their choice of research design to be adopted for their clinical studies. One of the most important ways for a clinical researcher to gain a better understanding of the epidemiology of a disease is to develop a clinical data registry for that disease and then conduct a registry-based study. The WHO defines a patient registry as *"a file of documents containing uniform information about individual persons, collected in a systematic and comprehensive way, in order to serve a pre-determined scientific, clinical or policy purpose"* (Brooke, 1974).

On the other hand, The US National Committee on Vital and Health Statistics defines a patient registry as *"an organized system for the collection, storage, retrieval, analysis, and dissemination of information on individual persons who have either a particular disease, a condition (e.g., a risk factor) that predisposes them to the occurrence of a health-related event, or prior exposure to substances (or circumstances) which are known or suspected to cause adverse health effects"* (National Committee on Vital and Health Statistics, 2019). In its simplest form, a patient registry consists of a collection of records that were traditionally kept as hard copy files and stored in cupboards; however, nowadays, all the data in these registries are usually compiled and then transformed into computerized data to be kept in a database.

A patient registry can adopt a different design depending on the objective for which the registry is created. For example, some registries require patients to be followed up for a period of time, usually years, in order to enable researchers to monitor the course of a disease, its treatment outcome and the survival of patients (Lim et al., 2008; Wan-Ahmad & Liew, 2016). On the other hand, there are other registries that are designed as a

cross-sectional study in which the data was gathered only once, of which the suicide registry is a good example (Ali et al., 2014).

Previously a patient registry usually aimed to focus mainly on collecting the administrative records of patients and was only collecting minimal data on clinical care. Nowadays, there is a growing impetus for advancement in knowledge of medical care via conducting useful and relevant clinical research, and along with the advent of modern-day technology. This has resulted in a tremendous increase in the variety and volume of data variables that are collected and stored in a patient registry (Pillay et al., 2008).

A patient registry is specifically designed to store all relevant data from a list of patients with a common medical condition. Since it contains a list of patients who share the same disease presentation with the same medical condition, therefore it can further be categorized by the disease (i.e. disease registry) or by the specific exposure to drug treatment (i.e. drug registry). Both these types of registries can prospectively collect a wide variety of information by using standardized questionnaires. Thus, both will have tremendous potential for enabling researchers who are working in the medical field to answer many important research questions pertaining to clinical practice. Studies based on registry data can often be used as a basis for decision-making purposes by providing a real-world view of various types of outcomes from clinical practices and healthcare services. Such purposes of registry-based research include quality improvement, benchmarking, clinical (or epidemiological) research, clinical effectiveness, cost-effectiveness, device surveillance, treatment surveillance and population surveillance.

Patient registry data can have a profound potential for answering dozens of research questions. However, the process of conducting a research study by using registry data can be very challenging. First and foremost, most of the time these registry data are not owned by the researchers. So, the researchers will need to formally request permission to gain access to

these registry data and also to use them for research purposes. Besides that, it is important for a researcher to obtain prior permission from the owners of registry data in order to ensure that the researcher will abide by all the terms and conditions applicable to gaining access to such registry data for research purposes, and also to be held responsible for maintaining the confidentiality of such data as long as they are still having physical custody of these data.

In addition, the whole procedure of data preparation can involve a challenging process since researchers will have to deal with many variables. There are several important procedures for the data cleaning process which are necessary for preparing a quality data set that is ready for analysis. Although all these are considered as the most basic procedures for conducting research using secondary data, however these must still be highlighted for research involving patient registry data as they deserve special attention in these registry-based studies due to the complex nature of their data collection.

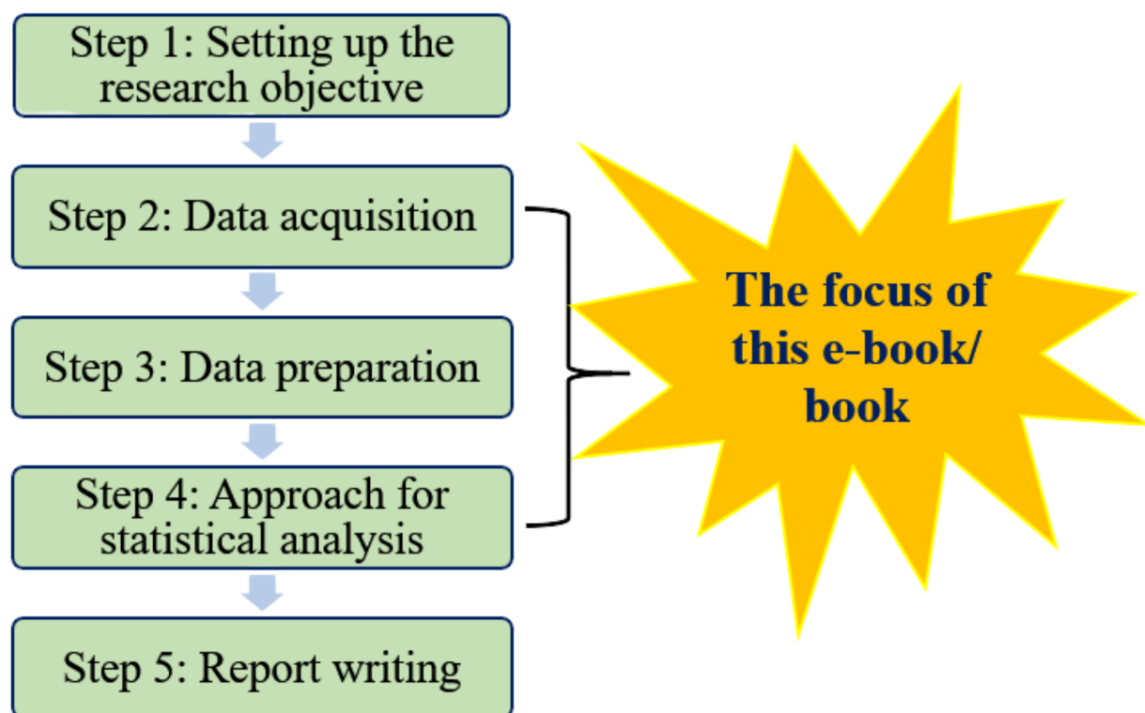


Figure 1.1: A common step-by-step process for conducting research using secondary data and the focus of this book

## 1.1 Purpose and Scope

Figure 1.1 shows a common step-by-step process to conduct research using secondary data, including data derived from a patient registry database. This overall process requires at least five important steps. As explained earlier, data retrieved from a patient registry database can potentially answer various important research questions and/or hypotheses, which in turn will determine the specific type and style of report writing to be prepared. Both these processes are highly dependent on the ability of the researchers in (i) determining the right research question to be answered and in (ii) determining how the discussion is going to be structured and presented.

In order to achieve both (i) and (ii), the researcher will need to have a prior in-depth understanding of the subject matter regarding the disease itself. Normally, the researcher should begin to acquire an in-depth understanding of the subject matter by reviewing the relevant literature, talking to the experts from the relevant field, and undertaking certain important observations. It will also be advantageous for the researcher if he/she can gain a first-hand experience in the subject matter itself by becoming the clinician who are is actively involved in the patient registry. This clearly shows that both the discussions on “Setting up the research objective” and “Report writing” are still very important. However, this e-book/book emphasizes the importance of these three processes such as “Data acquisition”, “Data preparation”, and “Approach for statistical analysis”, which are particularly relevant for a study being conducted by using patient registry data. A registry-based study cannot guarantee that its findings are valid and its report is well-written even if it has a valid research hypothesis, if the three processes mentioned above are not properly conducted. This is due to the specific nature and complexity of its data collection requirements.

Therefore, this e-book/book aims to provide a practical guide for conducting research by using data obtained from a patient registry database, which emphasizes the necessity to



balance the optimal and the feasible to maximize the gains within the constraints of the available resources of a registry database. Its practical significance is to render assistance to researchers in preparing the final data set obtained from a patient registry database for the purpose of subjecting them to a specific statistical analysis. To this end, the structure of this e-book/book is divided into three major sections, namely: (i) procedures of data acquisition, (ii) procedures for data preparation, and (iii) approach for statistical analysis.

## *Chapter 2 – Data Acquisition*

The first step in conducting research by using data from a patient registry database is to set up the right and valid research objective. It is the responsibility of the researchers to equip themselves with adequate prior knowledge of the subject matter in order to be able to formulate the right and valid research objective. Once the research objective is set, the next step is to retrieve the relevant data from a patient registry database. However, before retrieving and analysing any patient registry data, it is first necessary to determine whether it is truly feasible to retrieve the necessary data from a patient registry and also to analyse them in order to address specific clinical research questions. Hence, the following description provides a list of steps to take when planning to determine the feasibility of data collection for a patient registry.

### **2.1 Obtain Prior Consent and Study Approval**

The initial set-up of most registries is usually proposed by either the local government (most likely by the Ministry of Health), or a group of specialists in which the registries are formed as part of a non-government organization, NGO's initiative (Lim et al., 2008; Ali et al., 2014; Wan-Ahmad & Liew, 2016). All clinical data obtained from each individual patient are considered as confidential data. Hence, it is necessary to obtain prior consent from the respective authorities for using these registry data for research purposes (including publishing the results of any studies which are being conducted by using these registry data). After consent has been successfully sought from all the relevant authorities, both parties (i.e. data owner and researcher) will then need to agree upon to adhere to a set of data sharing procedures that will prevent any potential breach(es) of patient confidentiality.

Besides seeking consent from the respective data owner, it is also necessary to obtain study approval from the authorized ethics committee. These are common practices for most

of the countries in the world today. Although the study does not involve exposing any intervention(s) to the patients and/or it does not even involve any interaction with the patients, however ethics approval is still necessary to ensure that privacy and confidentiality of all patient data are adequately protected in order that such data will not be abused. In addition, most reputable journals will require ethical approval as a prerequisite for publication.

Once the consent and the approval have been granted, the data can be transferred from one party to the other in several ways, including by hand or by an external gadget (such as the use of a CD or thumb drive), or via sharing through a password-protected cloud storage. It is always advisable to take special precautions for handling all the patient data which are collected by a registry because they are strictly confidential. Hence, both the data owner and the researcher must adhere to a proper code of practice for data sharing procedures which will duly ensure that confidentiality of all these patient data is strictly maintained at all times. Apart from this, the terms of reference must also be carefully drafted between the both parties so that they have been made aware of each other's rights and responsibilities after having agreed to work together to accomplish a shared goal. It will also be ideal to carefully document the terms of reference between the two parties for future reference.

## **2.2 Understand the Scope of a Registry**

After having obtained prior consent and approval from all relevant authorities, the next step is to understand the scope of data collection necessary to meet research needs for a patient registry and the minimum standard of data quality necessary for fulfilling all the purposes of the registry. Therefore, it is of utmost importance for the researcher to gain a thorough understanding of the background of subject matter which underpins the content of a registry by reading around the subject matter, its proposed method of data collection and all

terminologies adopted by the registry. This can be achieved by gleaning relevant information from the instruction manuals of other similar registries and also from any previous publications of the same registry.

One of the ways to understand the scope of a registry is by reviewing its Case Report Form (CRF). The CRF is a structured questionnaire designed to collect information from the patients. It is a questionnaire where specific details of each individual patient including his/her profile, history of disease, treatment and outcomes are recorded (Please see Appendix 2). The researcher should also carefully examine the design of a CRF, including the definitions of all its variables, in order to assess whether the data collected by the CRF meet the minimum data requirement set by the researchers. It is also strongly recommended for both the CRF and its accompanying instruction manual to be kept for future reference. If the researcher has any doubts about the data collection procedures of a registry and/or the data collected by its CRF, it would be in his/her best interest to seek for a clarification with the person in-charge of data management for the registry data at the earliest opportunity.

Next, the researcher should determine whether the data collected by the registry will be able to address all primary and secondary research objectives of the study. This will involve a careful assessment of whether the registry is collecting all the necessary data which are required by the researchers and also whether the sample size deployed by the study will be adequate for fulfilling all research objectives of the study. The researcher should also carefully screen through the CRF to confirm if all the variables that will be collected by the CRF are actually reported in the registry database.

Notwithstanding the above, it is always possible for some of the variables found in the registry not to be collected by all participating centres, which could possibly due to a lack of comparable, standardized data entry mechanism (i.e. pending data entry) or sometimes due to a lack of IT support from the staff at the participating centre. This will result in the presence

of missing data in the registry, which the researcher will have to decide whether to wait for a future data installation, or to proactively fill in these missing data, or to impute a median value for all the missing data, or to simply declare them as missing data if the minimum sample size deployed by the study is already adequate.

### **2.3 Estimate the Minimum Sample Size Required for a Registry Study**

Sample size calculation is an important consideration because all researchers have to enrol a minimum number of patients or to retain enough patients in the study in order to attain the desired power of a registry study. Generally, it will not be necessary to perform a sample size calculation if the registry is created using a census approach. However, many problems can potentially arise when missing data are found in the registry. If this happens, then it will be necessary to estimate the minimum sample size required for the registry-based study even if the registry has been created by using the census or population-based approach. In a similar vein, it is also always necessary for the researcher to give due consideration to the sample size calculation of a registry-based study if the registry has been created by collecting a sample of the total population data.

A notably significant observation that universally applies to all clinical research is that sample size required for a study will always depend on the study objective. For example, when calculating the sample size required for an objective of estimating the prevalence rate or the population mean, the values of alpha, power of the study and margin of error are three important determinants which must be set beforehand. Based on the findings provided by Cochran's (1977), the sample size required to estimate the prevalence rate within a population will be 384 if the researcher has set the values of alpha, power and margin error at 0.05, 80% and 5.0% respectively (Krejcie & Morgan, 1970). Since the values of alpha and power are always set at 0.05 and 80.0% respectively and the value of margin of error of 5% is generally

acceptable (i.e. neither too large nor too small), the sample size of 384 is generally considered to be sufficient for estimating the prevalence rate for most of the cases.

In contrast to prevalence rate, the sample size required to estimate the population mean is usually smaller (Bujang et al., 2012; Bujang et al., 2015). However, this may not hold true if the margin of error selected by the researcher is different, since a smaller margin of error will always necessitate a bigger sample size. Therefore, it is up to the researcher to decide how small the margin of error in sample populations they intend to detect, since a larger sample will be required to detect a smaller margin of error.

For those studies that involve hypothesis testing, the sample size required is determined by three components, such as type I error (usually fixed at 0.05), power (usually fixed at 80.0%) and the effect size. The effect size is estimated by using a different formula, which shall depend on the specific statistical test employed. It is recommended that a sample size of at least 500 to be required for performing most of the common statistical hypotheses (Bujang et al., 2015). Nevertheless, researchers should always bear in mind that all sample size calculations will have to take into account all the specific study objectives, in order to ensure that the power attained by the study will be sufficient to address all the study objectives.

There are numerous guidelines that currently exist in the literature for estimating the minimum sample sizes required for performing (i) the sensitivity and specificity test, (ii) the correlation test, (iii) the intra-class correlation coefficient, (iv) kappa agreement, (v) multiple linear regression and (vi) analysis of covariance, (vii) Cronbach's alpha test, (viii) logistic regression, (ix) survival analysis (Concato et al., 1995; Bujang et al., 2016; Bujang & Baharum, 2016; Bujang & Baharum, 2017a; Bujang & Baharum, 2017b; Bujang et al., 2017; Bujang et al., 2018a; Bujang et al., 2018b).

## **2.4 Prepare a Dummy Table for Statistical Analysis**

After having obtained consent from all the relevant authorities, and also assessed the feasibility of establishing the registry (i.e. by ensuring a sufficiently high quality of data collected by an adequate sample size of the registry data), then the subsequent step is to prepare a dummy table which displays the overall format and structure of layout of the results to be presented. Dummy tables are actually empty tables containing only the variable items along with their statistical measures, and which will only be filled by actual data after data analysis has been performed (see Figure 2.1).

Using the analogy of building a house – constructing a dummy table for a data set is similar to drawing a schematic plan for the house. In other words, the dummy table shall display the probable final output of a data set which is designed to answer the objective(s) of the study. A dummy table can be designed as long as all the variable items and their units of measurement, along with the specific statistical analyses to be conducted on the variables are known (for example: to measure age in years and then to calculate its mean and its standard deviation).

Researchers will first have to decide the appropriate statistical measures to be reported and then to select the best format to present the results in a most informative way. An important way to get some idea of how to prepare a dummy table is to adopt the overall structure and format of dummy tables which commonly appear in previous research publications, which are often found in standard research reports or scientific journals. In addition, there are also some standard ways of designing a dummy table for presenting both descriptive and inferential statistics which have already been proposed by scholars (Lang & Secic, 1997).

A good dummy table serves several important functions, namely : (i) to confirm that the stated hypothesis/objective is doable and achievable, (ii) to provide a template for

performing the analysis systematically, (iii) for documentation purposes (for example: a dummy table can be one of the stipulated requirements of a research proposal), (iv) to facilitate an effective two-way communication between the researcher and the statistician, and (v) to enable the researcher to keep his/her focus on data collection and data analysis (so that the initial presentation of data can serve as a useful starting point for a discussion between researcher and statistician to enable both parties to mutually understand and agree on the research question and/or hypothesis, which can lead to a further refinement of researcher's intentions).

Each dummy table should be accompanied by a proper title. This title should be very specific, and it can also be stand-alone. For instance, an example of a title for a dummy table that describes demographic profile of patients in a registry dataset can be written as follows: "Table 1: Demographic profile of patients with type 2 diabetes mellitus enrolled in National Diabetes Registry from 1<sup>st</sup> January until 31<sup>st</sup> December 2018 in Malaysia." Hence, we can see that a proper title for a dummy table should contain the variables/domains (for example: demographic profile of patients), study population (for example: patients with type 2 diabetes mellitus), name of registry (for example: National Diabetes Registry, Malaysia) and the duration of the registry (for example: from 1<sup>st</sup> January until 31<sup>st</sup> December 2018).

This will enable the readers to have a better grasp of the overall presentation of results in the dummy table to obtain a more specific understanding of the researcher's original intention (in order for the reader to avoid having misunderstood the true intention of the researcher). After having prepared the dummy tables, the researchers should carefully evaluate them to ensure that all the dummy tables are able to clearly depict all the variables in the study and also to present the results in a way that adequately addresses the study objectives. This will allow the researchers an opportunity to make any necessary final



amendments to these dummy tables if they do not adequately fulfill the study objectives and/or clearly present all the variables and results to the readers in a meaningful manner.

*Example of study results*

**Table 1.** Demographic profile of patients participated in the study (n=1332)

Profile	Mean (SD)	n	%
<b>Age</b>	54.4 (15.4)	54.4	15.4
<b>Gender</b>			
Male		679	51.0
Female		653	49.0
<b>Race</b>			
Malay		675	50.7
Chinese		491	36.9
Indian		149	11.2
Others		17	1.3
<b>Education level</b>			
No formal education		103	7.9
Primary		347	26.5
Secondary		625	47.8
Tertiary		232	17.8

*Thus, here is the dummy table*

**Table 1.** Demographic profile of patients participated in the study (n=1332)

Profile	Mean (SD)	n	%
<b>Age</b>			
<b>Gender</b>			
Male			
Female			
<b>Race</b>			
Malay			
Chinese			
Indian			
Others			
<b>Education level</b>			
No formal education			
Primary			
Secondary			
Tertiary			

Figure 2.1: An illustration of a template dummy table versus an actual dummy table filled up with study results

## *Chapter 3 – Data Preparation*

One of the compulsory requirements for conducting a quantitative research is data preparation. There is a wide variety of many different types of data which will be collected by a patient registry. The data retrieved from the registry should have been formatted in a way which makes them amenable to subsequent statistical analysis. Thus, the data analysis may not make sense if proper and adequate data preparation procedures have not been implemented.

The complexity of the data preparation procedures for patient registry data will depend on many factors, including the sheer volume of registry data, the intended scope of the study, its study design, study objectives and many more. Hence, each of these factors will impose its specific requirements on the data preparation of registry data, which is necessary for transforming these data into a form that can produce meaningful results upon statistical analysis.

Therefore, after all the data have been collected or derived, it is very important for the researchers to prepare a proper data preparation plan, which shall incorporate several important considerations such as (i) handling of duplicates found in a data set, (ii) matching and combining data from different data sets when necessary, (iii) specifying the plans for data analysis that correspond to major aims and objectives of a registry-based study, (iv) data cleaning, (v) handling missing data in a registry.

		id	gender	height	weight		
		134	1	1.6	77.0		
		156	1	1.7	128.0		
		225	1	1.5	74.0		
		188	2		78.5		
Duplicate		321	2	1.8	53.2		Missing value
		111	2	1.9	100.0		
		156	1	1.7	201.0		Extreme value
		269	2	1.3	77.0		
		166	1	1.5	128.0		
		116	2	1.4	74.0		
Inconsistency		117	3	1.7	78.5		
		119	1	1.7	79.9		
		200	1	1.3	80.2		
		210	1	1.8	100.0		
<u>DUPLICATES</u>							
More than one observation having same unique <i>patient ID</i>							
<u>MISSING VALUES</u>							
Blank cell without data or information							
<u>INCONSISTENCY</u>							
Data that is not consistent with planned code							
(i.e. Male=1 and Female = 2, thus code 3 is inconsistent)							
<u>EXTREME VALUES</u>							
Value that had exceeded the plausible and logical lower & upper limit							

Figure 3.1: Visual example and definition for duplicates, missing values, inconsistency and extreme values

### **3.1 Proper Handling of Duplicates Found in a Dataset**

A common problem which often arises from a patient registry database is the presence of duplicate patient data (see Figure 3.1). The ideal data-capture mechanism for a registry is to allow each unique individual to be enrolled in a disease registry only once: for example, the National Diabetes Registry (NDR) allows the input of an individual patient's data only once (National Diabetes Registry, 2013). However, duplicates can possibly occur if the same patient had been registered in several different clinics or hospitals, which makes it difficult to track the same patient across multiple systems to identify those patients who are found to have been duplicated between various systems. It is an important and necessary step to remove duplicate patient data because they will lead to an overestimation of both the incidence and prevalence of disease. The final results will also be biased if there are too many duplicates because they can cause the actual population or census data to deviate from the truth.

To remove duplicate patient data in the registry, a researcher must take the first step in identifying those variables within the registry which are used to distinguish each individual patient, often referred to as the patients' unique identifiers. These identifiers commonly include a patient's name, his/her identity card (IC) number or identification number. In addition, the researcher will also need to pre-specify the criteria for determining a case as a duplicate. For example: would it be acceptable for those patients who had been enrolled more than once in the same centre to be regarded as duplicates? In the case of the Malaysian National Diabetic Registry, the answer is 'yes' (National Diabetes Registry, 2013).

However, it has been found that different patient registries may operate differently by imposing different criteria for the identification of duplicates. For instance, the National Eye Database (NED) allow the same patient to be enrolled in the same registry twice if and only if he/she were having problems in both eyes. In this case, the patient's unique identifier will

offer a better method for patient identification by indicating his/her patient identification number and also specifying which eye he/she is having problems (Mohamad-Azaz & Goh, 2018).

After having compiled a list of criteria to identify a case as a duplicate, the researcher will then need to adopt a systematic approach for removing the duplicates. Using NDR report as an example to illustrate the situation in which a patient may have been enrolled in more than one centre on two different dates, the investigator may decide to retain only one copy of a patient's medical record in the registry, by basing either on time (for example: to keep the one prepared on the earliest or most recent date) or (ii) by basing on the content of the medical record (for example: to keep the one containing more information).

For the former, the selection can be made very easily (and probably even quicker) by using an IT software. However, it may pose a risk of losing vital patient information if the selected patient's medical record is not the one having the most complete set of information pertaining to the patient's medical condition. In contrast, selecting the record containing the most complete set of information will be very time-consuming since it will require a very careful evaluation of each medical record before omitting them one at a time. After removing all the duplicates, it is necessary to perform a validation step to ensure there are no more duplicates to be found in the data set.

An efficient way to avoid duplicate data being entered into the registry is to create a mechanism in the system in which a duplicate will be detected as soon as the same patient's identifiers (such as his/her name or identity card number) have been entered into the registry more than once. By using this approach, the database system will keep a permanent record of a patient's data whenever his/her data have been entered into the registry database, which will then alert the investigator as soon as the data from the same patient are being entered into the registry again.

This implementation of a set of proper validation steps for detection of duplicates can save the investigator a lot of time for removing duplicates. However, even with the implementation of such proper validation steps for detection of duplicates, it is still compulsory for the investigator to carefully screen for duplicates in the data set before conducting the statistical analysis because we cannot be totally sure that the system would be working as efficiently as we have expected them to.

Table 3.1: Summary table for FAQs on handling duplicates

No.	Questions	Answers
1	What is duplicates?	When more than one observation or subject sharing one unique identifier.
2	Does duplicate data always need to be removed?	First of all, researchers will need to clearly define what is considered as a unique observation. Ideally, each unique patient should have only one record in the same registry. However, certain registries such as registry for eye problem may allow the same patient to be enrolled in the same registry twice if and only if he/she were having problems in both eyes. In this case, the patient's unique identifier will then indicate his/her patient identification number and also specify which eye he/she is having problems.
3	What can invalid duplicates do to your data?	Results can be rendered biased and invalid.
4	How to remove invalid duplicates? Can duplicates be avoided in the future?	The easiest way is by using statistical software. The most important point to take note is that both researchers and statisticians will need to know the basis for detecting the duplicates. Besides that, researchers will need to have reached a consensus regarding the basis for deleting the duplicates in the database, which should be standardized across all the dataset.
5	Can duplicates be avoided in the future?	Yes, it is definitely possible to avoid having duplicate data in the registry database by improving the efficiency of the database so that the system will be able to detect and avoid allowing any invalid duplicate data to be entered into the database.



### **3.2 Match and Combine Data from Different Datasets**

It is not uncommon for data from a registry to be stored by various separate sections within a registry database. Therefore, it will be necessary to combine these sections together to form a single data set. For example, the National Cardiovascular Disease Registry (NCVD) has two main forms: namely the notification form and the follow-up form. Data obtained from both forms will be stored in two separate sub-databases. These two sub-databases will have to be merged during analysis since some of the results will require input from both sub-datasets (Wan-Ahmad & Liew, 2016).

Meanwhile, it is also possible for some of the study objectives to necessitate the establishment of a linkage between the registry data and the data from an external data set, such as matching the original registry data with the data obtained from the National Death Record to determine the survival rate of the patients (Wong & Goh, 2016). To link together the data obtained from two different data sets, both data sets will need to have unique identifiers which can be matched by either deterministic or probabilistic matching, which are two different strategies for record linkage or data matching. When the investigators are dealing with sensitive information such as patients' identifiers, they should ensure that prior consent has been obtained from the patients and all the necessary approvals have been granted by the respective authorities.

Before performing the matching process, the identifiers have to be distinctly unique in both data sets. Although matching can be performed by using statistical software, however it is still necessary to perform a validation step by conducting several random checks on these records to make sure that exact matching had been performed. This can be achieved by obtaining a random sample of several matched records and comparing them against the source data from the original data sets.

### **3.3 Set up Conditions and Requirements for Analysis**

Before subjecting the registry data to the proposed statistical analysis, the researchers will have to pre-specify the analytical principles and statistical techniques to be employed, such as the inclusion and exclusion criteria for data collection and analysis. One of the important pre-specified conditions for data analysis is the duration of the study period. For instance, the study period for a registry report that is published on a yearly basis will be from 1<sup>st</sup> January until 31<sup>st</sup> December of the same year (National Transplant Registry, 2015; Wong & Goh, 2016). When the research study or its data analysis involves only a subset of the patients found in a registry, then it is necessary to clearly specify a list of strict inclusion criteria for subject selection. For example, although the diabetes registry includes both type 1 and type 2 diabetic patients, however the researcher may intend to study them separately (and therefore both types of patients will be analysed and reported separately) (Bujang et al., 2018c; Bujang et al., 2018d).

### **3.4 Establish Data Cleaning Procedures**

Now, we have pre-specified the plans for data analysis of the data set. The next step we shall take is to perform data cleaning. At this stage, all the relevant variables will be evaluated to determine whether they are properly coded and labelled, all the values are expressed within the pre-specified range, and the layout of all relevant data have been properly organized for the subsequent analysis. Some of the most common indicators of good quality data are (i) all duplicates have already been removed from the data set, (ii) all the outliers which exist only due to incorrect or invalid data entry have already been removed from the data set (i.e. the truly valid observations should be kept even though they resemble the outliers), (iii) all data inconsistencies have been rectified or (for example, sex is male and pregnant status is 'yes') and all missing values have been fill in or imputed or declared as missing.

It is also a good practice for the researcher to pre-specify an acceptable range for all numerical variables. This is important because it enables the researcher to detect any outliers, abnormal values and data inconsistencies during the data cleaning process. For example, say the acceptable range for an adult patient's age should be from 18 to 100 years. So, we should therefore query if a patient's age has been reported as 'more than 100 years'. For best practices in effective data management, the upper and lower limits of each numerical variable should be pre-determined and also clearly specified in the database system so that all out-of-range values will automatically be removed from the registries. In a similar vein to the procedure for removal of duplicates, it is also necessary to carefully screen for the presence of outliers and other abnormal values even though a proper set of data validation rules has already been established within the database system, as there is always a chance for a programming error to occur in the system.

There are two approaches to adopt for data cleaning, namely: (i) to clean the data obtained from its original source (front end), or (ii) to clean the data obtained from the database system (back end). Cleaning the data obtained from its original source is much more time-consuming because this will involve making reference to the original source documents or the subjects themselves in order to verify the accuracy of the data collected. Even though such a data cleaning process will be very likely to increase the validity of the data, however it may occasionally not be feasible due to both time and resource constraints.

An alternative approach is to apply statistical techniques for data cleaning (back end). These techniques commonly include data imputation or simply declaring the data as 'missing'. Undoubtedly, simply declaring the data as 'missing' is the easiest way out, which is usually applied whenever the proportion of missing values is very minimal. When preparing the registry data set for data analysis, it will first be necessary to create new variables and to

recode the existing variables which are commonly found in patient registry databases, just like those found in any other databases.

For example, a new variable called body mass index (BMI) can be derived from height and weight calculations, and then recoded into a set of varying categories according to a pre-specified classification system. While standardized classification systems are readily available for some variables such as BMI and stages of chronic kidney disease; many other variables do not have such standardized classification systems, and therefore it will be necessary to classify them by using other means.

For instance, age can be categorized into either a '5-year' or '10-year' age groups. However, since there is usually a widely-accepted way of categorizing a variable, the researcher may review the existing literature to search for the most common way of categorizing a variable; which will then enable him/her to make a comparison of these variables between different studies. For example, the age group can be categorized according to an international standard which is widely used (WHO, 2007; WHO, 2008; Bujang et al., 2012). Once these new variables in a registry data set have been computed and/or recoded, these data should again be checked for accuracy in a validation step to ensure both the computation and the recoding of these data are correct.

Table 3.2: Summary table for FAQs on handling data cleaning

No.	Questions	Answers
1	What is data cleaning?	Is a process to handle missing values, eliminate invalid duplicates, outliers and inconsistency in the data. The aim of data cleaning is to produce a clean dataset that is ready for analysis.
2	What can poor data do to your results?	Results will be biased and invalid.
3	How to clean the data?	<p>Researchers and statistician need to determine the basis and condition to detect missing values, invalid duplicates, outliers and data inconsistency.</p> <p>When all the invalid data is detected, then necessary imputation or corrections need to be made.</p> <p>For researcher that are using registry data for research, data cleaning probably need to be done at back end which is when after data is obtained from the database system. To clean data from the front end (from its original source) usually will take longer time.</p>

### **3.5 Proper Handling of Missing Data in a Registry Database**

Missing data usually occur when no data is available for reporting a variable within the data set (see Figure 3.1). Missing data commonly occurs during the data collection process, especially for a patient registry database, and which can potentially lead to insufficient power for the registry study (due to an inadequate sample size resulting from missing data) (Kim & Curry, 1977). The presence of missing data can also render some common statistical analyses either invalid and/or unfeasible, and can also introduce a potential source of bias into the estimates derived from a statistical model (Rubin, 1987; Becker & Walstad, 1990). Therefore, it is necessary to handle these missing data by using appropriate analytical strategies to analyse the remaining set of incomplete data.

Missing data are categorized as either 'missing completely at random' (MCAR) or 'missing not at random' (MNAR) (Rubin, 1976). Missing data will be considered as MCAR when their occurrence is not influenced by other variables. For example, MCAR can happen when some questionnaires have been lost by accident, and also when the respondents have unintentionally overlooked some questions, or when the specimen container has been damaged by accident (which resulted in a loss of the results due to attrition of sample collected by the investigation). In these scenarios, the simplest technique for handling missing data will involve the use of ad hoc methods such as complete case analysis and available case analysis (pairwise deletion) in order to give unbiased results (Greenland & Finkle, 1995).

Unlike the MCAR data, the MNAR data are influenced by certain factors. For example, when asking a patient for his or her income level, the data may be more likely to be missing when the income is extremely high. The reason for this missing data is obviously unrelated to any visible patient characteristic. Another situation in which MNAR can also occur is when the resources are not available for a particular specimen collection. In these

scenarios, the researcher will be losing some vital information. Unfortunately, there is no universal method for handling such missing data (Rubin, 1987; Greenland & Finkle, 1995). Therefore, it is advisable to take proactive measures to avoid or at least minimize the chance of having such MNAR data in patient registries.

These missing data can also occur at any stage of the collection of registry data. Firstly, they can often arise because the required information has not been collected, which could possibly happen due to a variety of reasons such as the loss of a case record form (i.e. resulting in a MCAR) or the unavailability of resources (i.e. resulting in a MNAR). Secondly, it is also possible that the information has not been recorded on the case report form (CRF), even though it is easily available. Thirdly, it is also possible that although the data are already available in the CRF, the procedure of data entry has inadvertently omitted the recording of such data within the registry database. To understand how to deal with such missing data, the researcher should determine at which stage the data have gone missing in order to decide the right course of action to take. Since most researchers are keen to proceed with data analysis as soon as possible, they may adopt any one of the three possible ways to handle these missing data, namely (i) by using a validation technique, (ii) by using an imputation technique and the last and also the easiest way will be to (iii) simply use a specific code to designate the missing data as 'missing'.

The validation technique for handling missing data refers to a method whereby the missing data are being substituted by using relevant supplementary information obtained by the other variables in the registry data set. Using the Malaysian identification card number as an example, it is possible for a researcher to realistically determine the date of birth (by referring to the first five digits), the gender (i.e. an odd last digit indicates male and an even last digit indicates female) and the province/state in which the patient was born (by referring

to the sixth and seventh digits' number) by referring to the 12-digit number on the person's identification card.

Although it is always possible to have erred by determining someone's ethnicity by basing on his/her name alone, the above-mentioned method is considered a more objective way to determine a person's ethnicity, which makes it more relevant for the Malaysian patients. Apart from this, it is usually possible to decipher a person's gender based on his/her name alone. However, the researcher must also have to accept the fact it is also possible for a deviation from the truth to occur if he/she is imputing a 'missing' gender by basing on the person's name alone. Thus, it is necessary to declare such limitations in the study report or manuscript in order to make them transparent to all the readers.

On the other hand, imputation is a technique which replaces the missing data with a probable value which has been estimated by mathematical computations. There are many imputation techniques which can be applied in patient registry databases, such as the single imputation methods (including mean imputation, median imputation, conditional mean imputation) and the multiple imputation methods (including hot decking method and multiple imputation by chained equations). However, a full discussion of all these imputation techniques is beyond the scope of this paper, and so the reader is advised to look up other relevant sources for further information about these techniques.

The decision to apply imputation techniques to rectify the error caused by missing data also depends on the objective of the study. In some technical reports, it is necessary to report missing data in order to evaluate the competency of the registry for collecting information. Hence, imputation will not be necessary in such instances. When imputation is deemed necessary in other circumstances, then the researcher shall have to select an appropriate imputation technique and to provide a valid justification for it.



The easiest way to handle missing data is to simply declare them as missing. A unique code will be introduced to represent the missing values across all the variables. Assuming the researcher uses the code "0" to define missing, and at the same time the same code "0" is also used in another variable to represent answer "No" (for example: "1" for yes and "0" for no); then another distinctly different code such as "9999" will be a better option to represent the missing value instead of "0". In some instances, it is possible for several different codes to represent the missing values, albeit each with a different definition. For example, code "9999" can be used to represent the true missing value, whereas code "8888" can be used to indicate that the variable is not relevant or not applicable to a particular patient (such as the 'pregnant' status for a male patient). Hence, it becomes necessary to adopt a different approach for the analysis of these two different types of missing values.

Irrespective of how missing data will be dealt with, they can always be easily detected during the data cleaning process. Simple descriptive analysis such as the percent frequency (%) will be able to detect the total number of missing data that are found in a registry database. Then, the researcher shall need to identify to which individual patient the missing values actually belong by basing them on the individual identifiers such as the patient identification number. Finally, the researchers will need to obtain a consensus among themselves on the most appropriate way to handle the missing values. After having identified the best way of handling the missing values, these missing values can then be replaced accordingly by using an appropriate imputation technique. In addition, a full description of the way in which the missing data are being replaced shall also be provided in the study report or manuscript, in order to ensure that the selection of any imputation techniques that have been applied for replacing the missing data are fully justified by the researchers and are also made clear and transparent to the reader.

Table 3.3: Summary table for FAQs on handling missing data

No.	Questions	Answers
1	What is missing data?	All empty observations in any of variables in the dataset
2	What missing data can do to your result?	For large volume of missing data, results derived from the analysis can be biased and invalid.
3	How to handle missing data?	<p>First of all, researcher should determine at which stage the data have gone missing in order to decide the right course of action to take. Since most researchers are keen to proceed with data analysis as soon as possible, they may adopt any one of the three possible ways to handle these missing data, namely</p> <ul style="list-style-type: none"> <li>• by using a validation technique (eg: Using the Malaysian identification card number to determine the date of birth (by referring to the first five digits),</li> <li>• by using an imputation technique and</li> <li>• simply use a specific code to designate the missing data as 'missing'.</li> </ul>

### 3.6 A Summary of Data Preparation Procedures

The aim of data cleaning is to produce a data set that is usable and ready to be analysed. Thus, the data set should be free from invalid duplicates, missing values, inconsistent values and extreme values (or outliers). Below is an example of a problematic data set as presented in Figure 3.2. Can you spot what are the problems with this data set?

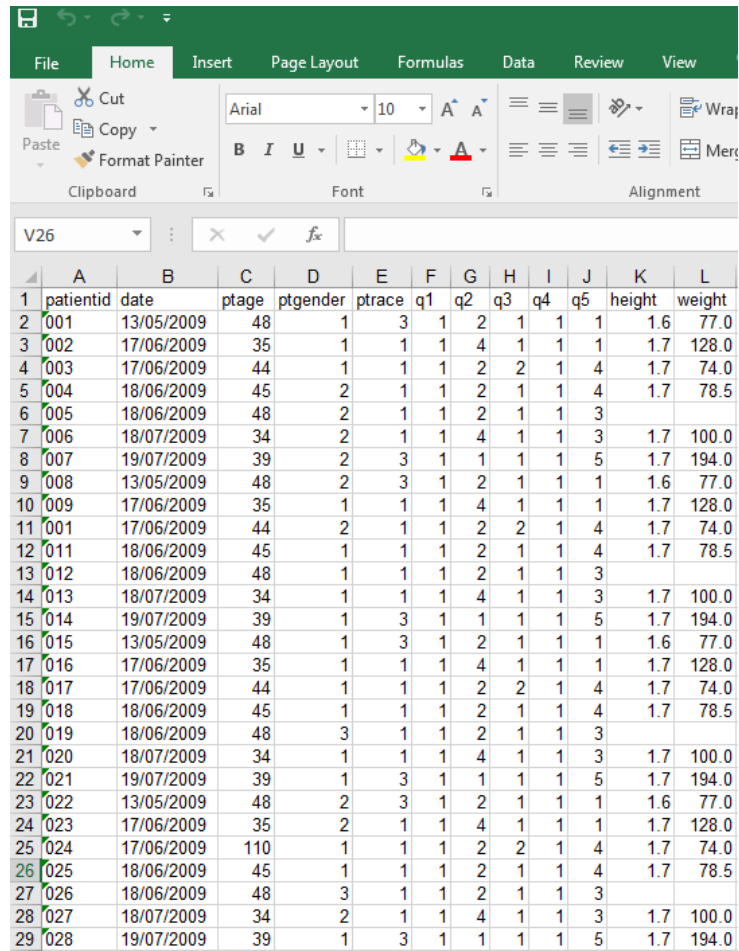
id	gender	age	race	smoke	weight	bsa	pulse	stress	abp_cat	Pregnant
1	female	44	malay	yes	85	1.75	63	33	normal	1
2	female	50	chinese	yes	94	2.1	70	14	high	1
3	3	50	chinese	yes	95	1.98	72	10	high	1
4	male	51	indian	yes	95	2.01	73	60	high	0
5	female	49	.	no	89	1.89	72	55	normal	1
6	female	58	malay	yes	100	2.25	71	90	high	1
7	male	52	malay	no	100	2.25	69	42	high	0
8	male	46	malay	no	91	1.9	66	18	normal	0
9	male	50	chinese	no	89	1.83	69	62	normal	0
10	female	48	chinese	no	93	2.07	64	35	normal	1
11	male	54	indian	yes	94	2.07	74	90	high	0
12	female	50	chinese	no	94	1.98	71	21	normal	1
13	3	51	indian	no	92	2.05	68	47	normal	1
14	female	445	malay	yes	87	1.92	67	80	normal	1
15	male	45	malay	no	101	2.19	76	98	normal	0
16	female	48	chinese	no	95	1.98	69	65	normal	1
17	female	42	malay	yes	87	1.87	62	18	normal	1
18	female	47	chinese	no	95	1.9	70	12	high	1
19	male	47	malay	no	91	1.88	71	99	normal	0
20	male	57	indian	yes	96	2.09	75	99	high	1

Figure 3.2: Sample of problematic data

The problems are as followed:

1. Inconsistent code for gender (id=3 and id=13)
2. Patient's age of 445 years (id=14)
3. Missing value for race (id=5)
4. Male patient with a 'pregnant' status (id=20)

Perhaps it is easy to eyeball these problems since the data set is small. Large data sets with dozens of variables are usually complicated and will require highly skilled statisticians (who will be using an appropriate software) to handle the data. To solve these problems, a statistician will need to verify the accuracy of these data by referring to the original records, to rectify any errors found in these data or for certain variables such as age, to apply an appropriate imputation technique for replacing any missing values in the data set. After data cleaning has been completed, it is necessary to keep this 'clean' data set by saving it in a different file. A good data set should contain the right number of subjects with all their related variables lumped together for each category of these variable definitions (see Figure 3.3 and Figure 3.4). Finally, it is also necessary to give a proper name for the data set in order to facilitate future retrieval, such as "osa\_study20090930" which indicates a project related to Obstructive Sleep Apnoea (OSA) study locked at 30<sup>th</sup> September 2009.



	A	B	C	D	E	F	G	H	I	J	K	L
1	patientid	date	ptage	ptgender	ptrace	q1	q2	q3	q4	q5	height	weight
2	001	13/05/2009	48	1	3	1	2	1	1	1	1.6	77.0
3	002	17/06/2009	35	1	1	1	4	1	1	1	1.7	128.0
4	003	17/06/2009	44	1	1	1	2	2	1	4	1.7	74.0
5	004	18/06/2009	45	2	1	1	2	1	1	4	1.7	78.5
6	005	18/06/2009	48	2	1	1	2	1	1	3		
7	006	18/07/2009	34	2	1	1	4	1	1	3	1.7	100.0
8	007	19/07/2009	39	2	3	1	1	1	1	5	1.7	194.0
9	008	13/05/2009	48	2	3	1	2	1	1	1	1.6	77.0
10	009	17/06/2009	35	1	1	1	4	1	1	1	1.7	128.0
11	001	17/06/2009	44	2	1	1	2	2	1	4	1.7	74.0
12	011	18/06/2009	45	1	1	1	2	1	1	4	1.7	78.5
13	012	18/06/2009	48	1	1	1	2	1	1	3		
14	013	18/07/2009	34	1	1	1	4	1	1	3	1.7	100.0
15	014	19/07/2009	39	1	3	1	1	1	1	5	1.7	194.0
16	015	13/05/2009	48	1	3	1	2	1	1	1	1.6	77.0
17	016	17/06/2009	35	1	1	1	4	1	1	1	1.7	128.0
18	017	17/06/2009	44	1	1	1	2	2	1	4	1.7	74.0
19	018	18/06/2009	45	1	1	1	2	1	1	4	1.7	78.5
20	019	18/06/2009	48	3	1	1	2	1	1	3		
21	020	18/07/2009	34	1	1	1	4	1	1	3	1.7	100.0
22	021	19/07/2009	39	1	3	1	1	1	1	5	1.7	194.0
23	022	13/05/2009	48	2	3	1	2	1	1	1	1.6	77.0
24	023	17/06/2009	35	2	1	1	4	1	1	1	1.7	128.0
25	024	17/06/2009	110	1	1	1	2	2	1	4	1.7	74.0
26	025	18/06/2009	45	1	1	1	2	1	1	4	1.7	78.5
27	026	18/06/2009	48	3	1	1	2	1	1	3		
28	027	18/07/2009	34	2	1	1	4	1	1	3	1.7	100.0
29	028	19/07/2009	39	1	3	1	1	1	1	5	1.7	194.0

Figure 3.3: A sample of data set in excel sheet based on Obstructive Sleep Apnea (OSA) study

Data Dictionary			
Column	Variable	Description	Condition/Labelling
1	patientid	patient id	001 - 235
2	date	date of notification (day/month/year)	formatted as xx/xx/2009
3	ptage	age in years	numerical
4	ptgender	patients gender	1 - Male and 2 - Female
5	ptrace	patients ethnicity	1 - Malay, 2 - Chinese, 3 - Indian and 4 - Others
6	q1	Do you snore?	1-Yes, 2-No, 3-Do not know
7	q2	Snoring loudness	1)Loud as breathing 2)Loud as talking 3)Louder than talking 4)Very loud
8	q3	Snoring frequency	1) Almost every day 2) 3-4 times/week 3) 1-2 times/week 4) 1-2 times/month
9	q4	Does your snoring bother other people?	1-Yes, 2-No, 3-Do not know
10	q5	How often have your breathing pauses been noticed?	1) Almost every day 2) 3-4 times/week 3) 1-2 times/week 4) 1-2 times/month
11	height	height in cm	set at one decimal point
12	weight	weight in kg	set at one decimal point

Figure 3.4: Variable definition of a sample of data set in Excel sheet based on Obstructive Sleep Apnea (OSA) study (as shown in Figure 3.3)

## *Chapter 4 – Approach for Statistical Analysis*

It is important to always bear in mind that prior to subjecting the registry data to the proposed statistical analysis, researchers must be clear about the intended study objectives and the significance of all the study variables in relation to the objectives of the study. All statistical analyses should be conducted based on the research question and the study objectives. These analyses should be conducted only after data cleaning has been performed and the data are properly organized for the subsequent statistical analyses. Data cleaning for a patient registry database can often be a very difficult task because of the sheer volume and the wide variety of data that can possibly be collected by a registry.

Therefore, the function of data preparation of patient registry data (which includes data collection, data preparation and data analysis) is usually performed by experienced statisticians or their equivalents. Say, for example, a researcher or data analyst is in charge of conducting the statistical analysis of patient registry data; it is then likely for him/her to be mentored and supervised by a senior statistician or a statistical consultant. In addition, the researcher or data analyst should also be well-versed in the handling of patient registry data, and be equipped with adequate prior knowledge in the design of these registries (including their clinical parameters, definitions of all the variables), the design of its case report form, the procedures for data collection and data preparation for statistical analysis of all the patient registry data, all of which have already been discussed previously.

#### **4.1 A Proposed Guide for Statistical Analysis of Patient Registry Data**

Depending on the research question and the study objectives, the statistical analysis of the registry data can range from a simple analysis such as the calculation of percentage frequency and other descriptive statistical analyses to more advanced statistical analysis such as multivariate regression and disease modelling. The following is a compilation of all the recommended steps for planning a robust way to conduct statistical analysis of patient registry data.

##### *Before statistical analysis*

- a. Data must be free from invalid duplicates, inconsistency and outliers before beginning to conduct any of the proposed statistical analyses. Only outliers which occur due to incorrect data entry must be removed from the data set [i.e. this means that the true observations (albeit they resemble outliers) should be retained for subsequent statistical analysis.]
- b. Researchers should be fully equipped with prior knowledge on the design of these registries (including their clinical parameters, definitions of all the variables and their relevant terminologies), the design of its case report form and all the data collection procedures.
- c. The proposed statistical analysis should itself align with the research question, the study objectives and the layout of the data in the dummy tables.
- d. Researchers must have the full understanding of how the estimation of the required sample size of registry data has been specifically tailored for the subsequent statistical analysis of registry data, as this data set shall become the analysis set for the study.



*During statistical analysis*

- e. For a descriptive study (using census data) with no missing values, it is still acceptable to provide just a descriptive statistical analysis of the registry data alone, which means that no inferential statistical analysis will be conducted.
- f. However, if there are missing values in the data set, then it will become necessary to conduct both descriptive and inferential statistical analysis of the data set since it will now be necessary to infer these findings about the target population from the sample data.
- g. It is important to observe and take proactive steps to ensure that a list of underlying assumptions which must hold true in order for all the statistical computations to be valid, especially with regard to parametric test and regression modelling. This is because it will be a futile attempt to conduct statistical analysis by performing all statistical computations and yet violating these underlying statistical assumptions (or failing to ensure that these assumptions actually hold true), as this will yield invalid conclusions.
- h. Conducting inferential statistical analysis on patient registry data (such as the use of disease modelling methods as research tools) can be a very complicated task because these patient registries can have many dozens of variables which are regarded as the independent variables for a measurable outcome. It will not be possible to label all these variables as risk factors in a single analysis by using a multivariate model because such an analysis will not be efficient due to issues which have arisen from not being able to (i) fulfil all the assumptions in regression modelling and also to (ii) meet the minimum sample size requirements. Therefore, it is strongly recommended for the researcher to carefully select a list of independent variables that qualify to be regarded as risk factors to be tested in the

subsequent analysis, and such a selection which shall be based on the underlying judgment about the scientific relevance, scope and significance of the study.

- i. Some of the variables may have smaller sample sizes. Researchers may have to decide whether these variables should rightfully be regarded as the independent variables (i.e. the factors associated with an outcome) or the dependent variables (i.e. the outcome) for the subsequent statistical analysis, as this decision will affect their sample size requirements, especially for multivariate analysis.
- j. It is also strongly recommended conducting these data analyses by using the programming codes because it promotes transparency. Besides that, a detailed step-by-step mechanism of the analysis (and together with the flow diagram of its procedures) should also be clearly documented for future reference if necessary.
- k. Avoid analysing any variable which is not related to the study objective because it is necessary to conduct statistical analysis according to a well-thought-out data analysis plan which has been formulated to address hypotheses and aims of the research, and also to account for possible confounders. Although it seems tempting for a statistician to analyse all the variables at first glance, due to the presence of so many different variables in the entire data set; however, it is always recommended for a statistician to prepare beforehand a set of planned tables and figures (called dummy tables) which will provide a visual presentation of the layout of the results. This will avoid confusion among both researcher and statistician because such dummy tables will illustrate how the results will be displayed and also help to bring into focus of what both of them are doing. The researcher, on the other hand, will complement the statistician by carefully framing the research question and funnelling it down into testable hypotheses and action steps, which are detailed in a data analysis plan. This can realistically be

attained by the researcher since he/she is usually the subject-matter expert who has a sound understanding of the subject, and is therefore able to decide which variables will need to be analysed.

*After statistical analysis*

- l. For descriptive analysis, it is recommended that sample size (n) for each variable will also be reported, in order for the readers to obtain a rough estimate of the proportion of missing values.
- m. It is recommended to insert a footnote at the bottom of the page for alerting the readers that those findings obtained from the variables with a large number of missing values (say, the total number of data reported is less than 50% of the total population data) will have to be interpreted with caution.
- n. It is necessary to validate the results again by another senior statistician or statistical consultant. At the bare minimum, it will be necessary to run the same analysis at least twice, in order to maximize the chance for detection of errors.
- o. It is important to pre-specify any limitation(s) of the registry-based study pertaining to its achievement of research objectives that can possibly result from either the estimation of the variables found in the patient registry database, or from their sample size requirements, or from their proposed statistical analysis.

These are very useful tips which are highly recommended for a statistician to adopt when formulating an approach for statistical analysis of registry data. In order to serve as a quick guide to all statisticians and researchers alike, this e-book/book provided a summary checklist for the whole process of collecting and analysing registry data

including the approach to be adopted for statistical analysis of registry data (Please see Table 5.1).

## **4.2 Presentation of the Final Results**

The overall presentation of the final results should be formatted in such a way that will directly address the research objectives which were set earlier. Therefore, all tables and figures obtained from statistical analysis should be presented clearly in order to effectively convey its key message to its wide target audience, including all the stakeholders. There are various ways of presenting findings. The proper way for presenting the results shall depend on the research objective, scope of the proposed research study and type of target audience.

Most importantly, it is strongly recommended for both the statistician and researcher to have a clear idea of both the ideal and expected results, before commencing the statistical analysis. By doing so, the statistician can develop a versatile data analysis plan which will ideally be built on the research protocol. Having a clear plan of action for data analysis will also guard against data-driven results, which is important for research integrity and quality. Hence, the drafting of the dummy tables should be done collaboratively by both statistician and researcher, which is important for ensuring that the proposed data analysis will be feasible, specific and focused.

Since it can be a difficult task for a novice or inexperienced researcher to draft these dummy tables, it will be recommended for him/her to refer to other dummy tables found in previously published research articles, especially those from peer-reviewed journals, as a guide for them to follow when they are drafting a new dummy table. To illustrate this point, this e-book/book lists down a total of 33 published scientific articles found in peer-reviewed journals and were written by various authors, all of which had reported on research studies conducted using patient registry data (Please see Appendix 1).

After having presented the final results, the findings obtained from these results will have to be reviewed again by the investigators to ensure that they are scientifically sound and valid, logical, relevant, and be reflective of the total population. Although the statistical analysis is run by using a programming code, it is still possible for mistakes to occur. These mistakes can arise at any stage during the analysis; for example, from the data itself or during the process of analysis (such as the use of the wrong programming code).

To promote best practices for data analysis, the researchers are encouraged to carefully review and validate all the results obtained from statistical analysis. If possible, it is strongly recommended to validate the results obtained from all the tables or at least from a few tables chosen randomly to ensure that all the analyses are correct. In addition, it is also advisable to conduct an external validation check by a colleague(s), a senior statistician or the supervisor. Any necessary amendments and/or rectifications should be made immediately whenever errors have been found. This is an important validation step to ensure that an effective presentation of accurate and significant research findings can be delivered to key stakeholders and/or research clients later on.

#### **4.3 Development of the Statistical Analysis Plan**

Statistical Analysis Plan (SAP) is a document that records all the important steps for the overall framework of statistical analysis. This plan provides a documentation which describes the procedures for (i) handling duplicates, (ii) matching and combining data sets, (iii) setting the conditions and requirements for the statistical analysis, (iv) handling missing values and (v) developing the framework for the statistical analysis (which includes a detailed description of the all steps to be taken for the statistical analysis).

Hence, the SAP is usually prepared after the output obtained from the analysis has been finalised. This plan will then provide a documentary evidence of how the analysis has been prepared, conducted and presented. The programming codes that are used to produce the

results are usually regarded as strictly confidential but the SAP should be made available to the whole study team. This is necessary in order to ensure that all statistical analyses are made transparent to the stakeholders, which will enable them to understand exactly how all the statistical analysis procedures have been performed (so that they can pose any queries whenever necessary).

The SAP is not the same as the section of “statistical analysis” found in a full-text research paper commonly published in a journal. What is written in the statistical analysis section will only be a small part of the whole write-up of SAP, since the SAP is a written document which provides a complete and detailed description of all the specific step-by-step procedures for managing and conducting statistical analysis. Thus, the SAP can provide documentary evidence of the adoption of a valid approach for the entire process of both data preparation and data analysis. This SAP document should be retained for future reference and in some cases, it may be necessary for it to be made available in the future to various stakeholders as a documentary evidence of the use of robust statistical methods and procedures in producing the results (and also that all statistically analyses are conducted in a scientifically valid manner). This e-book/book has illustrated the importance of a SAP by providing a checklist which highlights a list of key elements of a properly-prepared and complete SAP (Please see Table 5.1).

The efficient way of conducting the statistical analysis is to develop a programming code for each analysis. The codes can be reused many times to produce the updated results whenever a new data set becomes available. This will increase the efficiency of the analysis. Another advantage of using the programming software is that the codes can be regarded as a documentary evidence of whether the selection of the subjects and the analysis were conducted in a correct manner by referring to each of the lines inside the programming codes. Therefore, all the steps that are outlined in the programming codes for conducting the study

should be made transparent to the research team and the key stakeholders and also be clearly documented in Statistical Analysis Plan (SAP).

Since the generation of output from these patient registry reports is regarded as a repetitive work, therefore it is necessary to select an appropriate statistical software. This software should incorporate the use of programming codes in the analysis. The recommended statistical software is STATA (Copyright 1996–2018 StataCorp LLC), SAS (Copyright © 2011, SAS Institute Inc., Cary, NC, USA) and R (R Foundation for Statistical Computing, Vienna, Austria). Besides that, it is also necessary to ensure that the computer has fulfilled certain competency requirements, such as the requisite standards for a sufficient RAM and also appropriate working processor, in order to ensure that it has the capacity to conduct statistical analysis on an extremely large data set.

If the analysis is not generated by using the programming codes, then it is important to carefully document all the necessary input processes and outputs obtained from the analysis in a structured work diary for our future reference. Regardless of whether the analysis has been generated by programming codes or not, all the input processes (for example: handling duplicates, matching and combine datasets, setting the conditions for analysis, data cleaning and management, handling missing values and developing an overall framework for the analysis) should be fully documented in the Statistical Analysis Plan, which shall serve as a useful future reference.

## *Chapter 5 – Summary*

This e-book/book has sought to provide a brief guide on how to conduct research by using data obtained from a patient registry database. It can be a difficult task to deal with patient's registry data because it revolves around major ethical issues pertaining to data privacy and confidentiality issues (since these are sensitive data about each individual patients). In addition, there are numerous variables and each with tons of data in a patient registry that have rendered data handling much more difficult than that of the other types of survey research. Therefore, this e-book/book attempts to provide an overview of the framework for all the steps necessary to be taken by a researcher when conducting research using data obtained from a patient registry database. It shall also serve as a simple guideline which stipulates all the standard requirements that must be met for the purpose of obviating major problems from cropping up during the subsequent statistical analysis of these registry data.

The scope of this e-book/book is limited to discussion on data acquisition, data preparation and approach for statistical analysis only. Hence, this e-book/book will not be discussing the report writing which is always the final step in the research process. Since patient registry databases can potentially answer dozens of research questions, thus there are a myriad of different ways on how to discuss and present findings derived from the statistical analyses. As an illustration, this e-book/book has listed at least 33 scientific papers published in peer-reviewed journals about different types of research being conducted on data from patient registry databases.

All these studies had also involved numerous types of patient registries, including Adult Diabetes Control and Management (ADCM) Registry 2009 (currently known as National Diabetes Registry (NDR)), National Eye Database (NED), The Malaysian National Cardiovascular Disease Database (NCVD) registry, National Orthopaedic Registry Malaysia



(NORM), National Suicide Registry Malaysia (NSRM), National Database on Children and Adolescent with Diabetes (DiCARE), and many more. All the registries aim to recruit patients at a national level, which are initiated by hospitals and Ministry of Health Malaysia. (Please see Appendix 1).

Finally, this book concludes with a summary which provides a recommended step-by-step guide of the flow of this procedure (as presented in Figure 5.1) and also a structured checklist (as presented in Table 5.1) that can be very useful for guiding and facilitating a statistician for conducting research by using data obtained from a patient registry database.

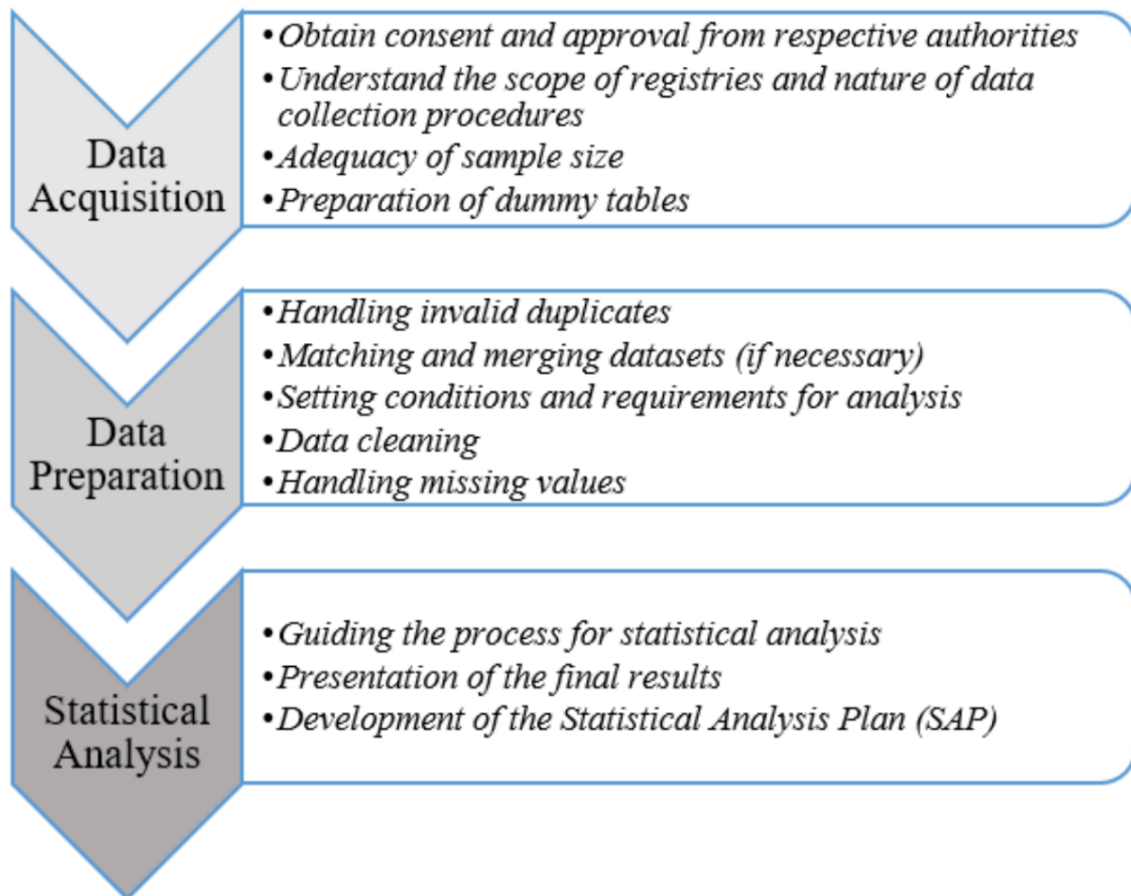


Figure 5.1: A practical guide consisting of a list of recommendations for planning a research study by using data from patient registries

Table 5.1: Checklist for data acquisition, data preparation and approach for statistical analysis for research using patients' registries databases

No.	Section and topic	Item	State 'Yes' or 'No'
<b>1</b>	<b>Data acquisition</b>		
<i>i.</i>	<i>To obtain consent and study approval.</i>		
i (a)	To ensure a valid consent has been obtained from all relevant authorities for accessing the data for research purposes.	( )	
i (b)	To obtain ethical approval from relevant authority(ies)	( )	
i (c)	To ensure a secure method of data transfer is in place to ensure both privacy and confidentiality of data is maintained.	( )	
<i>ii.</i>	<i>To understand both the scope of registry and nature of data collection procedures.</i>		
ii (a)	To review all data elements in the case report form (CRF)/questionnaire and the definitions of all the outcome and exposure variables are included in the CRF.	( )	
ii (b)	To discuss the detailed data collection procedures with the subject-matter experts to ensure that all terminologies have been correctly defined, and the exact step-by-step data collection procedures are feasible and accurate.	( )	
ii (c)	To achieve a thorough understanding of the background of the subject matter by reading relevant previously-published scientific reports or research manuscripts.	( )	
<i>iii.</i>	<i>To determine a minimum sample size required for the patient registry study.</i>		
iii (a)	To ensure that all the variables of interest as stated in the CRF/questionnaire (for both outcome and exposure variables) are available in the data set.	( )	
iii (b)	To ensure that a sufficiently large sample size has been obtained for each of the variables of interest to address all the research objectives of the study.	( )	
<i>iv.</i>	<i>To prepare dummy tables for illustrating the layout of data in the patient registry study report.</i>		

iv (a)	To gain a full understanding of the research question, the broad scope of the study and also all its research objectives.	( )
iv (b)	To retrieve a dummy table that is relevant to this study by searching for such dummy tables which are found in published scientific reports and/or research manuscripts.	( )
iv (c)	To prepare the dummy tables for their subsequent presentation to key stakeholders and/or research clients in order to obtain a consensus of opinion for adopting such dummy tables in the layout of data presentation within the registry study report.	( )
iv (d)	To decide an appropriate study title for each dummy table.	( )

## 2 Data preparation

### *i. To establish proper procedures for handling duplicates found in a data set.*

i (a)	To reach a consensus among the team of investigators for the definition for 'duplicates', and then to decide how to identify and handle them.	( )
i (b)	To validate the processes and/or measures to be taken for ensuring that the process of removing duplicates (and also for cleaning the data set) are adequate for ensuring that all data are correct, consistent and usable.	( )
i (c)	To record all the amendments that have been made in handling the duplicates.	( )

### *ii. To match and merge data obtained from different data sets (if necessary).*

ii (a)	To obtain consent from all respective authorities for matching data obtained from external sources.	( )
ii (b)	To ensure that each record is labelled with a unique individual identifier before performing the matching process.	( )
ii (c)	To validate the merged records by ensuring that the data sets have been accurately merged.	( )
ii (d)	To record all the amendments that have been made in merging the data sets (if necessary).	( )

### *iii. To set up conditions and requirements for analysis*

iii (a)	To ensure that all the data have been arranged according to their planned start date and end date.	( )
iii (b)	To pre-determine a list of criteria for selecting subjects who are included in the registry study, such as inclusion and exclusion criteria.	( )

iii (c)	To pre-determine the criteria for the data set ( ) undergoing further analysis (if necessary).
iii (d)	To record all the setting conditions and ( ) requirements for analysis.
<i>iv. To perform data cleaning/ quality check</i>	
iv (a)	To ensure that all the necessary variables have ( ) been correctly coded and labelled.
iv (b)	To ensure that there are no inappropriate duplicate ( ) cases are found in the data set.
iv (c)	To ensure that there are no invalid outliers or any ( ) other invalid 'extreme' values found in the data set.
iv (d)	To ensure that there is no data inconsistency found ( ) in the data set.
iv (e)	To ensure that all new variables that are derived ( ) from other variables have been validated (for example, body mass index that was previously derived from weight and height will need to be validated again to ensure that all computations are correct).
iv (f)	To ensure that the data set has been organized in ( ) an appropriate format in order to prepare it for subsequent analysis.
iv (g)	To record all the amendments that have been ( ) made in performing data quality check.
<i>v. To have proper procedures for handling missing values.</i>	
v (a)	To ensure that a pre-specified and standardized ( ) code is used to represent missing values.
v (b)	To ensure that the replacement of missing values ( ) by using an appropriate imputation method for certain variables will be justifiable and appropriate for this purpose.
v (c)	To describe the imputation method in the ( ) Statistical Analysis Plan (SAP) if it has been used for filling in the missing data.
v (d)	To record all the amendments that have been ( ) made in handling missing values.

### **3 Approach for statistical analysis**

<i>i. To guide the preparation of the final data set for statistical analysis.</i>	
i (a)	To ensure that the process of data cleaning and ( ) data presentation are completed.
i (b)	To ensure that researchers are equipped with ( ) sufficient knowledge of the registry data including

	the variables, the terminologies used in the CRF and also all the data collection procedures.	
i (c)	To provide proposed statistical analysis of the registry data based on its research objectives and also on the presentation of data.	( )
i (d)	To report the minimum sample size (n) required for each variable so that the investigators can estimate the proportion of missing values.	( )
i (e)	To ensure adequate measures are being taken to validate the results obtained from the study (i.e including the need for repeating statistical analysis and re-checking the fulfilment of all underlying statistical assumptions).	( )
i (f)	To pre-specify the use of statistical software for performing the statistical analysis.	( )
i (g)	To ensure that all the steps of statistical analysis are properly documented which are based on programming codes, or all the details of the steps of statistical analysis are properly recorded in a structured work diary or in another standard format.	( )
<i>ii. To present the results obtained from the statistical analysis.</i>		
ii (a)	To ensure that all results are consistent with study objectives.	( )
ii (b)	To ensure that all the tables and figures are presented in an acceptable manner.	( )
ii (c)	To ensure that all the results have been carefully reviewed by another colleague(s) or by a senior statistician or by his/her supervisor before presenting them to key stakeholders and/or research clients.	( )
ii (d)	To ensure that all researchers/stakeholders have an opportunity to review and validate all the results obtained from the study.	( )
ii (e)	To ensure that all necessary amendments and corrections are being made in the results prior to the presentation of these results to key stakeholders and/or research clients.	( )
ii (f)	To ensure that all the results have been verified by key stakeholders and/or research clients.	( )
<i>iii. To write-up the statistical analysis plan.</i>		
iii (a)	To state the start date and end date for the dataset derived from the registry.	( )
iii (b)	To state number of records available within the two periods (start date and end date).	( )



iii (c)	To ensure that a proper procedure for handling duplicates has been described in the statistical analysis plan.	(     )
iii (d)	To ensure that a proper procedure for combining different data sets has been described (if necessary).	(     )
iii (e)	To ensure that a proper procedure for data preparation has been described.	(     )
iii (f)	To ensure that a proper procedure for deriving new variables from an existing registry data set has been described (if necessary).	(     )
iii (g)	To specify all the conditions for statistical analysis such as the criteria for selecting the subject for subsequent analysis.	(     )
iii (h)	To ensure that a proper procedure for handling missing data has been described.	(     )
iii (i)	To state the number of sample size in each analysis set.	(     )
iii (j)	To ensure that all statistical techniques that have been applied in the statistical analysis are fully reported.	(     )
iii (k)	To ensure that the type of statistical software that is used for the analysis is reported in the final study report.	(     )

## References

Ali N, Zainun K, Haniff J, et al. (2014). Pattern of suicides in 2009: data from the National Suicide Registry Malaysia. *Asia Pac Psychiatry*, 6, 217–225.

Becker WE, Walstad WB. (1990). Data loss from pretest to posttest as a sample selection problem, *The Review of Economics and Statistics*, 72(1), 184–188.

Brooke EM. (1974). The current and future use of registers in health information systems, World Health Organization, Geneva.

Bujang MA, Ghani PA, Zolkepali NA, Selvarajah S, Haniff J. (2012). A comparison between convenience sampling versus systematic sampling in getting the true parameter in a population: explore from a clinical database: the Audit Diabetes Control Management (ADCM) registry in 2009. *IEEE Int Conf Stat Sci Bus Eng*.

Bujang MA, Abdul-Hamid AM, Zolkepali NA, Hamedon NM, Mat-Lazim SS, Haniff J. (2012). Mortality rates by specific age group and gender in Malaysia: Trend of 16 years, 1995–2010. *J Health Inform Dev Ctries*, 6(2), 521–529.

Bujang MA, Sa'at N, Joys AR, Ali MM. (2015). An audit of the statistics and the comparison with the parameter in the population. *AIP Conference Proceedings*, 1682, 050019. doi: 10.1063/1.4932510

Bujang MA, Adnan TH. (2016). Requirements for Minimum Sample Size for Sensitivity and Specificity Analysis. *J Clin Diagn Res*, 10(10), YE01–YE06.



Bujang MA, Baharum NA. (2016). Sample size guideline for correlation analysis. *World Journal of Social Science Research*, 3(1), 37–46.

Bujang MA, Baharum N. (2017a). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. *Arch Orof Sci*, 12(1), 1–11.

Bujang MA, Baharum N. (2017b). Guidelines of the minimum sample size requirements for Kappa agreement test. *Epidemiol. Biostat. Public Heal.* 2017;14(2). doi:10.2427/12267.

Bujang MA, Sa'at N, Tg-Abu-Bakar-Sidik TMI. (2017). Determination of minimum sample size requirement for multiple linear regression and analysis of covariance based on experimental and non-experimental studies. *Epidemiol. Biostat. Public Heal*, 14(3). doi:10.2427/12117.

Bujang MA, Omar ED, Baharum NA. (2018a). A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. *Malays J Med Sci*, 25(6), 85–99.

Bujang MA, Sa'at N, Tg-Abu-Bakar-Sidik TMI, Lim CJ. (2018b). Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malays J Med Sci*, 25(4), 122–130.

Bujang MA, Kuan PX, Tiong XT, et al. (2018c) The All-Cause Mortality and a Screening Tool to Determine High-Risk Patients among Prevalent Type 2 Diabetes Mellitus Patients. J Diabetes Res, Article ID 4638327.

Bujang MA, Tiong XT, Saperi FE et al. (2018d). The all-cause mortality and risk factors for mortality within five years among prevalent Type 1 Diabetes Mellitus Patients. Int J Diabetes Dev Ctries. doi:10.1007/s13410-018-0686-2.

Concato J, Peduzzi P, Holford TR, et al. (1995). Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy, J Clin Epidemiol, 48:1495–501.

Greenland S, Finkle WD. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol, 142, 1255e64.

Kim JO, Curry J. (1977). The treatment of missing data in multivariate analysis. Sociological Methods and Research, 6(2), 215–240.

Krejcie RV, Morgan DW. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607–610.

Lang TA, Secic M. (1997). How to Report Statistics in Medicine, Pa: American College of Physicians, Philadelphia.

Lim TO, Goh A, Lim YN, Morad, Z. (2008). Review article: use of renal registry data for research, health-care planning and quality improvement: what can we learn from registry data in the Asia-Pacific region? *Nephrology*, 13(8), 745–752.

Mohamad-Aziz S, Goh PP (Eds). (2018). Tenth Report of the National Eye Database 2016', National Eye Registry, Ministry of Health, Kuala Lumpur.

National Committee on Vital and Health Statistics. Frequently Asked Questions about Medical and Public Health Registries. Available at <http://ncvhs.hhs.gov/9701138b.htm>. (Accessed April 2019).

National Diabetes Registry. National Diabetes Registry (NDR) Report, 2009-2012. (2013). Non-Communicable Disease Section Disease Control Division Department of Public Health, Ministry of Health Malaysia, Kuala Lumpur.

National Transplant Registry (NTR). (2015). Twelfth Report of the National Transplant Registry, Kuala Lumpur, 2015

Pillay MS, Noor-Hisham A, Zaki M, Lim TO, Jamaiah H, Jaya-Purany SP. (2008). Patient registries in Malaysia and the role of the Clinical Research Centre of the Ministry of Health. *Med J Malaysia*, 63(Suppl C), 1–4.

Rubin DB. (1987). Multiple Imputation for Nonresponses in Surveys, John Wiley & Sons, New York.

Rubin DB. (1976). Inferences and missing data. *Biometrika*, 63, 581e90.

Wan-Ahmad WA, Liew HB. (Eds). (2016). Annual Report of the NCVD-PCI Registry, Year 2013-2014, National Cardiovascular Disease Database, Kuala Lumpur.

WHO. (2007). Bulletin of the World Health Organization. Development of a WHO growth reference for school-aged children and adolescents. World Health Organization, Geneva.

WHO. (2008). Causes of child deaths under 5. Summary of data and methods used. World Health Organization, Geneva.

Wong HS, BL Goh (Eds). (2018). Twenty Forth Report of the Malaysian Dialysis and Transplant 2016, The National Renal Registry, Kuala Lumpur.

## Appendix 1

### Examples list of scientific articles published based on data from patient registries by the authors

#### *Year 2011 and before*

1. Fuziah MZ, Hong JYH, Zanariah H, Harun F, Chan S P, Rokiah P, Wu LL, Rahmah R, Jamaiah H, GeetaA, Chen W S, Adam B. A National database on children and adolescent with diabetes (e-DiCARE): Result from April 2006 to June 2007. *Med J Malaysia*. 2008;63(Supplement C):37-40.
2. Chew BH, Ismail M, Shariff-Ghazali S, Lee PY, Cheong AT, Ahmad Z, Taher SW, Haniff J, Mustapha FI, Bujang MA. Determinants of uncontrolled hypertension in adult type 2 diabetes mellitus: an analysis of the Malaysian diabetes registry 2009," *Cardiovascular Diabetology*. 2012;11(1);54.
3. Chew BH, Ismail M, Lee PY, Taher SW, Haniff J, Mustapha FI, Bujang MA. Determinants of uncontrolled dyslipidaemia among adult type 2 diabetes in Malaysia: The Malaysian Diabetes Registry 2009. *Diabetes Research and Clinical Practice*. 2012; 96(3):339–347.

#### *Year 2012*

4. Ali NH, Zainun KA, Bahar N, Haniff J, Abd-Hamid AM, Bujang MA, Mahmood MS, NSRM study group. Pattern of Suicide in 2009: Data from The National Suicide Registry Malaysia. *Asia-Pacific Psychiatry*. 2012;6(2):217-225.
5. Ponniah JP, Shamsul AS, Adam BM. Predictors of mortality in patients with Acute Coronary Syndrome (ACS) undergoing Percutaneous Coronary Intervention (PCI): Insights from National Cardiovascular Disease Database (NCVD), Malaysia. *Med J Malaysia*. 2012;67(6):601-605.

#### *Year 2013*

6. Lee PY, Cheong AT, Ahmad Z, Ismail M, Chew BH, Ghazali SS, Bujang MA, Syed-Alwi SAR, Haniff J, Taher SW. Does Ethnicity Contribute to the Control of Cardiovascular Risk Factors Among Patients with Type 2 Diabetes? *Asia Pac J Public Health* July 2013;25(4):316-325
7. Chew BH, Shariff-Ghazali S, Ismail M, Haniff J, Bujang MA. Age  $\geq 60$  years was an independent risk factor for diabetes-related complications despite good control of cardiovascular risk factors in patients with Type 2 Diabetes Mellitus. *Experimental Gerontology*. 2013;48(5):485–491.
8. Chew BH, Shariff-Ghazali S, P Y Lee, A T Cheong, I Mastura, J Haniff, M A Bujang, S W Taher, F I Mustapha. Type 2 Diabetes Mellitus Patient Profiles, Diseases Control and Complications at Four Public Health Facilities - A Cross-sectional Study based on the Adult Diabetes Control and Management (ADCM) Registry 2009. *Med J Malaysia*. 2013;68(5):397-404.
9. Cheong AT, Lee PY, Shariff-Ghazali S, Bujang MA, Chew BH, Ismail M, Haniff J., Syed-Alwi SAR, Taher SW, Mat-Nasir N. Poor glycemic control in younger women attending Malaysian public primary care clinics: findings from adults' diabetes control and management registry. *BMC Family Practice* 2013;14:188.

10. Chew BH, Ismail M, Bujang MA. Comparing the disease profiles of adult patients with type 2 diabetes mellitus attending four public health care facilities in Malaysia. *Malaysian Family Physician* 2013;8(3):11-18.
11. Ahmad WA, Ali RM, Khanom M, Han CK, Bang LH, Yip AF, Ghazi AM, Ismail O, Zambahari R, Hian SK. The journey of Malaysian NCVD-PCI (National Cardiovascular Disease Database-Percutaneous Coronary Intervention) Registry: a summary of three years report. *Int J Cardiol*. 2013;165:161–164.
12. Lee CY, Hairi NN, Wan Ahmad WA, Ismail O, Liew HB, Zambahari R, et al. Are There Gender Differences in Coronary Artery Disease? The Malaysian National Cardiovascular Disease Database – Percutaneous Coronary Intervention (NCVD-PCI) Registry. *PLoS ONE*. 2013;8(8):e72382.

#### *Year 2014*

13. Shariff-Ghazali S, Ismail M, Ahmad Z, Cheong AT, Bujang MA, Haniff J, Lee PY, Syed-Alwi SAR, Chew BH, Taher SW. Control of glycemia and other cardiovascular disease risk factors in older adults with type 2 diabetes mellitus: Data from the Adult Diabetes Control and Management, Geriatrics & Gerontology International 2014;14(1):130-137.
14. Ali NH, Zainun KA, Bahar N, Haniff J, Abd-Hamid AM, Bujang MA, Mahmood MS. Pattern of suicides in 2009: Data from the National Suicide Registry Malaysia. *Asia-Pacific Psychiatry*. 2014;6(2):217–225.
15. Lee MY, Goh PP, Salowi MA, Adnan TH, Ismail M. The Malaysian Cataract Surgery Registry: cataract surgery practice pattern. *Asia Pac J Ophthalmol (Phila)*. 2014; 3: 343–347.

#### *Year 2015*

16. Bahar N, Wan-Ismail WS, Hussain N, Haniff J, Bujang MA, Abd-Hamid AM, Yusuff Y, Nordin N, Ali NH. Suicide among the youth in Malaysia: What do we know? *Asia-Pacific Psychiatry*, 2015;7(2):223-229.
17. Shariff-Ghazali S, Ismail M, Cheong AT, Bujang MA, Jamaiah H, Lee PY, Syed-Alwi SAR, Chew BH. Predictors of poor glycaemic control in older persons with type 2 diabetes mellitus. *Singapore Med J*. 2015;56(5):284-290.
18. Chew BH, Lee PY, Cheong AT, Ismail M, Bujang MA, Haniff J, Taher SW, Goh PP. Complication profiles and their associated factors in Malaysian adult type 2 diabetes mellitus—an analysis of ADCM registry. *International Journal of Diabetes in Developing Countries*. 2015;35(2):1-12
19. Mohd-Kasim K, Bujang MA, Abdullah MAH, Pan CH, Johari J, Stanley-Ponniah JP, Ibrahim S, Ang HL. Characteristics and outcome of patients with hand and upper limb infection in diabetes patients. *International Journal of Diabetes in Developing Countries*. 2015;35(2):123-128.
20. Salowi MA, Goh PP, Lee MY, Adnan TH, Ismail M. The Malaysian Cataract Surgery Registry: profile of patients presenting for cataract surgery. *Asia Pac J Ophthalmol (Phila)*. 2015; 4: 191–196.

#### *Year 2016*

21. Johari J, Bujang MA, Abdullah MAH, Mohd-Kasim K, Stanley-Ponniah JP, Ibrahim S, Pan CH. Pattern of Organisms and Antibiotics Used in Treating Diabetes Foot Infection. *International Medical Journal Malaysia*. 2016;15(1):25-30
22. Chew CH, Woon YL, Amin F, Adnan TH, Abdul-Wahab AH, Zul Edzhar Ahmad, Bujang MA, Abdul-Hamid AM, Jamal R, Chen WS, Hor CP, Yeap L, Hoo LP, Goh PP, Lim TO. Rural-urban comparisons of dengue seroprevalence in Malaysia. *BMC Public Health*, 2016;16:824.

#### *Year 2017*

23. Ganeshan M, Bujang MA, Soelar SA, Karalasingam SD, Suharjono H, Jeganathan R. Importance of Adopting BMI Classifications Using Public Health Action Points to Delineate Obstetric Risk Factors Resulting in Worsening Obstetric Outcomes Among Asian Population. *The Journal of Obstetrics and Gynecology of India*. 2018;68:173-178
24. Bujang MA, Adnan TH, Hashim NH, Mohan K, Ang KL, Ahmad G, Haniff J. Forecasting the Incidence and Prevalence of Patients with End-Stage Renal Disease in Malaysia up to the Year 2040. *International journal of nephrology*. 2017; Article ID 2735296: 5 pages.
25. Salowi MA, Chew FLM, Adnan TH, King C, Ismail M, Goh PP. The Malaysian Cataract Surgery Registry: risk Indicators for posterior capsular rupture. *Br J Ophthalmol* 2017;101:1466–1470.
26. Ho SF, Adnan TH, Goh PP. Prevalence and factors associated with second eye cataract surgery and the trend in the time interval between the two eye surgeries based on the Malaysian National Eye Database. *Asia Pac J Ophthalmol (Phila)*. 2017;6(4):310–317.
27. Lee MY, Adnan TH, Mariam Ismail M, Goh PP. The Malaysian Cataract Surgery Registry: Surgically induced astigmatism in phacoemulsification cataract surgery. *Asian J Ophthalmol*. 2017;15:159-171.

#### *Year 2018*

28. Bujang MA, Kuan PX, Tiong XT, et al. The All-Cause Mortality and a Screening Tool to Determine High-Risk Patients among Prevalent Type 2 Diabetes Mellitus Patients. *Journal of Diabetes Research*. 2018, Article ID 4638327, 8 pages. <https://doi.org/10.1155/2018/4638327>
29. Abdullah AAS, Ismail I, Bujang MA. Association of risk factors with a major re-amputation in Malaysian diabetic patients: a retrospective cohort analysis of patient registry. *Int J Diabetes Dev Ctries*. 2018;38:95.
30. Wai YZ, Fiona Chew LM, Mohamad AS, et al. The Malaysian cataract surgery registry: incidence and risk factors of postoperative infectious endophthalmitis over a 7-year period. *Int J Ophthalmol*. 2018;11(10):1685–90.

#### *Year 2019*

31. Bujang MA, Kuan PX, Sapri FE, Liu WJ, Musa R. Risk Factors for 3-Year-Mortality and a Tool to Screen Patient in Dialysis Population. *Indian J Nephrol*. 2019;29(4):235–241.
32. Bujang MA, Tiong XT, Saperi FE, Ismail M, Mustafa FI, Abd-Hamid AM. The all-cause mortality and risk factors for mortality within five years among prevalent Type 1 Diabetes Mellitus Patients. *Int J Diabetes Dev Ctries*. 2019;39:284-290.

*Year 2021*

33. Lee CY, Liu KT, Lu HT, Mohd Ali R, Fong AYY, Wan Ahmad WA (2021) Sex and gender differences in presentation, treatment and outcomes in acute coronary syndrome, a 10 year study from a multi-ethnic Asian population: The Malaysian National Cardiovascular Disease Database—Acute Coronary Syndrome (NCVD-ACS) registry. PLoS ONE 16(2): e0246474.



## Appendix 2

### Example part of a Case Report Form (CRF) from National Cardiovascular Disease Database (NCVD)

(Source: [http://www.acrm.org.my/ncvd/pciReport\\_15.php](http://www.acrm.org.my/ncvd/pciReport_15.php))

NATIONAL CARDIOVASCULAR DISEASE DATABASE (PCI REGISTRY) NOTIFICATION FORM				For NCVD Use only:	
Instruction: Complete this form to notify all PCI admissions at your centre to NCVD PCI Registry. Where check boxes <input type="checkbox"/> are provided, please check (✓) one or more boxes. Where radio buttons <input type="radio"/> are provided, check (✓) <u>only one</u> option.				Centre: <input type="text"/> ID: <input type="text"/>	
A. Date of Admission (dd/mm/yy): <input type="text"/>		B. Time of Admission (hh:mm): <input type="text"/> : <input type="text"/> (in 24hr clock)			
<b>SECTION 1: DEMOGRAPHICS</b>					
1. Patient Name: (as per MyKad / Other Document / ID)			2. Hospital RN :		
3. Identification Card Number:		Old IC No.			
MyKad: <input type="text"/>		<input type="text"/>			
Other ID Document No. <input type="text"/>		Specify type : (eg. passport, armed force ID) <input type="text"/>			
4. Gender: <input type="radio"/> Male <input type="radio"/> Female		5. Nationality: <input type="radio"/> Malaysian <input type="radio"/> Non Malaysian			
6a. Date of Birth: <input type="text"/> (write DOB as 01/01/yy if age is known)		6b. Age on admission: <input type="text"/> (auto calculate)			
7. Ethnic Group:		Foreigner, specify country of origin: <input type="text"/>			
<input type="radio"/> Malay <input type="radio"/> Punjabi <input type="radio"/> Melanau <input type="radio"/> Bidayuh <input type="radio"/> Chinese <input type="radio"/> Orang Asli <input type="radio"/> Murut <input type="radio"/> Iban <input type="radio"/> Indian <input type="radio"/> Kadazan Dusun <input type="radio"/> Bajau <input type="radio"/> Other Malaysian, specify: <input type="text"/>					
8. Contact Number: (1): <input type="text"/>		(2): <input type="text"/>			
<b>SECTION 2 : STATUS BEFORE EVENT</b>					
1. Smoking status: <input type="radio"/> Never <input type="radio"/> Former (quit >30 days) <input type="radio"/> Current (any tobacco use within last 30 days) <input type="radio"/> Not Available					
2. Medical history:					
a) Dyslipidaemia <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known		f) Documented Significant CAD <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known (Presence of >50 % stenosis on CTA, angiogram, ischaemia on functional cardiac imaging such as nuclear, MRI, echo or positive treadmill test. High calcium score alone is not sufficient)			
b) Hypertension <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known		g) New onset angina (<2 weeks) <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known			
c) Diabetes <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known		h) History of heart failure <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known			
<input type="checkbox"/> OHA <input type="checkbox"/> Insulin <input type="checkbox"/> Non pharmacology therapy/diet therapy		i) Cerebrovascular disease <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known			
d) Family history of premature cardiovascular disease (1st degree relative with either MI or stroke; <55 yold if Male & <65 yold if Female) <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known		j) Peripheral vascular disease <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known			
e) Myocardial infarction history <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known		k) Chronic renal failure (>200 µmol/L serum creatinine) <input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Not known On dialysis? <input type="radio"/> Yes <input type="radio"/> No			
<b>SECTION 3 : CLINICAL EXAMINATION and BASELINE INVESTIGATION</b>					
1. Anthropometric:		a. Height: <input type="text"/> (m) <input type="checkbox"/> Not Available		b. Weight: <input type="text"/> (kg) <input type="checkbox"/> Not Available	
2. Heart rate (at start of PCI): <input type="text"/> beats/min		3. Blood pressure (at start of PCI):		c. BMI: <input type="text"/> (auto calculate)	
4. Fasting Blood Glucose: <input type="text"/> mmol/L <input type="checkbox"/> Not Available		a. Systolic: <input type="text"/> (mmHg)		b. Diastolic: <input type="text"/> (mmHg)	
6a. Total cholesterol: <input type="text"/> mmol/L <input type="checkbox"/> Not Available		5. Hb A1c: <input type="text"/> % <input type="checkbox"/> Not Available		6b. LDL Levels: <input type="text"/> mmol/L <input type="checkbox"/> Not Available	
7. Baseline creatinine: <input type="text"/> µmol/L <input type="checkbox"/> Not Available		8. Baseline ECG:		<input type="checkbox"/> Sinus rhythm <input type="checkbox"/> 2 <sup>nd</sup> /3 <sup>rd</sup> AVB <input type="checkbox"/> RBBB <input type="checkbox"/> Atrial Fibrillation <input type="checkbox"/> LBBB <input type="checkbox"/> ST Deviation (for GRACE Score)	
9. Non Invasive Test:		i) <input type="radio"/> Done <input type="radio"/> Not Done		ii) Functional Ischaemia	
<input type="checkbox"/> Stress/ Exercise Test <input type="checkbox"/> Nuclear <input type="checkbox"/> MRI <input type="checkbox"/> Stress Echo <input type="checkbox"/> DSE <input type="checkbox"/> CT Scan				<input type="radio"/> Positive <input type="radio"/> Negative <input type="radio"/> Equivocal	
10. Glomerular Filtration Rate (GFR):		a. MDRD: <input type="text"/> mL/min/1.73m <sup>2</sup> (auto calculate)		b. Cockcroft-Gault: <input type="text"/> mL/min (auto calculate)	
Formula: GFR (Modification of Diet in Renal Disease (MDRD)) : $186 \times (\text{serum creatinine (micromol/L)})^{-1.154} \times (\text{age})^{-0.202} \times (0.742 \text{ if female})$ GFR (Cockcroft-Gault formula) : Male : $1.23 \times (140 - \text{Age}) \times \text{Weight (kg)} / \text{serum Creatinine (micromol/L)}$ Female : $1.04 \times (140 - \text{Age}) \times \text{Weight (kg)} / \text{serum Creatinine (micromol/L)}$					
<b>SECTION 4 : PREVIOUS INTERVENTIONS</b>					
1. Previous PCI:		2. Previous CABG:			
<input type="radio"/> Yes <input type="radio"/> No Date of most recent PCI (dd/mm/yy): <input type="text"/> <input type="checkbox"/> Not Available		<input type="radio"/> Yes <input type="radio"/> No Date of most recent CABG (dd/mm/yy): <input type="text"/> <input type="checkbox"/> Not Available			

## *About the authors*

### **Mohamad Adam Bujang**

Dr. Mohamad Adam Bujang is a Research Officer at the Institute for Clinical Research (ICR), Ministry of Health, Malaysia. Currently he is working as a biostatistician and researcher in the Clinical Research Centre, Sarawak General Hospital. He earned his PhD in Information Technology and Quantitative Sciences from MARA University of Technology in 2017.

### **Ang Swee Hung**

Dr. Ang Swee Hung is a medical officer at the Institute for Clinical Research (ICR), Ministry of Health, Malaysia. She started working in the research field since 2013, and was stationed at the Clinical Research Centre in Hospital Melaka before joining the headquarters of ICR in 2016. She completed her Master in Public Health in 2020 and is currently a candidate of Doctor in Public Health in University of Malaya, Malaysia.

### **Tg Mohd Ikhwan Tg Abu Bakar Sidik**

Mr. Tg Mohd Ikhwan Tg Abu Bakar Sidik is currently a PhD candidate in Universiti Kebangsaan Malaysia (UKM). Previously he was a biostatistician in the Institute for Clinical Research (ICR). He earned his Master in Statistics from Universiti Putra Malaysia in 2014.

### **Tassha Hilda Adnan**

Tassha Hilda Adnan is currently a Statistician in the Sector for Biostatistics & Data Repository, National Institutes of Health, Ministry of Health Malaysia. Formerly, she had joined Institute for Clinical Research, Ministry of Health from 2009 to 2019. She graduated with a Bachelor of Science (Hons.) Statistics from Universiti Teknologi MARA, Shah Alam in 2004.

### **Nadiah Sa'at**

Miss Nadiah binti Sa'at is a former Research Officer at the Institute for Clinical Research (ICR), Ministry of Health, Malaysia. She joined the Institute for Clinical Research, Ministry of Health in 2012, and currently she further her study in Medical Statistics at Universiti Sains Malaysia (USM). She earned her Bachelor in Mathematics from Universiti Putra Malaysia (UPM) 2007-2011.

**Hon Yoon Khee**

Mr. Hon Yoon Khee is a pharmacist working in the Institute for Clinical Research (ICR), Ministry of Health, Malaysia since May 2008 until now. He graduated with a Bachelor of Pharmacy degree from Otago University in New Zealand in December 1998, and had since been practising in community pharmacy and hospital pharmacy for several years before joining the Institute for Clinical Research, Ministry of Health in 2008. He also studied part-time clinical pharmacy from September 2007 until August 2012, and was awarded a Graduate Diploma in Clinical Pharmacy by University of South Australia in August 2012.

**Alan Fong Yean Yip**

Dr. Alan Fong Yean Yip is a Consultant Cardiologist in Heart Centre and a Network Head Clinical Research Centre, Sarawak General Hospital. He graduated from Royal College of Physicians, London, United Kingdom in 2004. He has various academic qualifications and the two latest were FRCP from Royal College of Physicians of Edinburgh in 2014 and FSCAI from Society of Cardiovascular Angiography and Interventions in 2017.



## *About the Publisher*

**Institute For Clinical Research,**  
National Institutes of Health (NIH), Ministry of Health Malaysia,  
Block B4, No.1, Jalan Setia Murni U13/52, Seksyen U13,  
40170 Shah Alam, Selangor Darul Ehsan, Malaysia.  
Phone: 603-3362 7700 | Fax: 603-3362 7701  
Email: [contact@crc.gov.my](mailto:contact@crc.gov.my)

*Analysis of patient data can be a complicated and challenging process, especially when the data involve many subjects and many variables. A patient registry is a database that organizes collecting the important set of data on a list of identifiable individuals for a specific disease. This type of data usually has tons of data and hundreds of different variables. Thus, the approach to conducting research by using a patient registry database will be more complicated than the other types of dataset. Since the handling of patient registry data is a challenging task, the authors have come out with this e-book/book to become a guideline for the statisticians, medical officers and scientists for them to refer as a handbook whenever they need to use patient registry data for their research.*

## **ABOUT THE AUTHOR**

***Dr. Mohamad Adam Bujang*** is a Research Officer at the Institute for Clinical Research (ICR), Ministry of Health, Malaysia. Currently he is working as a biostatistician and researcher in the Clinical Research Centre, Sarawak General Hospital. He earned his PhD in Information Technology and Quantitative Sciences from MARA University of Technology in 2017.

### **INSTITUTE FOR CLINICAL RESEARCH**

National Institutes of Health (NIH)  
Ministry of Health Malaysia  
Block B4, No.1,  
Jalan Setia Murni U13/52, Seksyen U13  
40170 Shah Alam, Selangor  
Malaysia  
Phone: 603-3362 7700  
Fax: 603-3362 7701  
Email: [contact@crc.gov.my](mailto:contact@crc.gov.my)