Routes-to-Diagnosis Analytics Standard Operating Procedure

Dr Finian Bannon and Dr Hannah McKenna, 2019

Queen's University Belfast

Contents

1. Introduction

1.1 Aim

The aim of this Standard Operating Procedure (SOP) is to describe the approach and methods adopted in the *Routes-to-Diagnosis* (RD) project to produce and present statistics in the final report. The SOP aims to allow an analyst to self-learn from the experience built up in the RD project, however a basic competence in Stata (a statistical package) is required. Under-lined terms in the SOP have an explanation in the Glossary (Section 6).

1.2 SOP Approach

This SOP is designed to be used in conjunction with the scripted syntax program files (*dofiles*) developed and executed in Stata (a statistical package) to implement the analytic methods. The *dofiles* contain all the programing that was employed in preparing the data and generating the statistical estimates. The *dofiles* have embedded comments (usually in green font behind an asterisk"*", or after two backward slashes) detailing what the following lines or blocks of programing code are designed to do; when the command's function is obvious, there may be no commentary. The *dofiles* can be reused for the next generation of RD estimates. The *dofiles* are also a record what and how tasks were carried out allowing for review and improvement. The SOP does not intend to replicate the *dofiles*, but orientate the analyst in how to use them, and provide more in-depth rationale, explanations, and references for methods than is possible in the comments in the *dofiles*.

1.3 Implementation of the analysis plan

A set of *dofiles* have been designed to carry out the RD analysis. A *dofile* contains Stata syntax code to carry out specific commands on the data, such as cleaning or analysis. A *dofile* can also call upon a second *dofile* to perform some commands required in the tasks of the first *dofile*. This second 'called' *dofile* often performs a frequent task requiring a set of commands. The entire set of tasks in the RD project can be executed by running a *dofile* called *MASTER.do,* which calls on secondary *dofiles* to complete the various components for the RD project. These secondary *dofiles* fall into data preparation (COMPILE DATASET.do), and analysis (PROPORTION.do, SURVIVAL.do, RANDOM EFFECTS MODELS.do) and presentation of results (BARCHART.do)).

1.4 General features of Stata coding in the *dofiles*

Maintaining tidy *dofiles*, with embedded comments, allows for easy review of the analysis that is carried out. Sections 1.4.1-1.4.3 describe some measures to eliminate repetitive code; this reduces the number of times that [repetitive] code needs to be changed as it is undergoing development.

1.4.1 Defined programs

Within a Stata session, a short program can be user-defined and submitted to the Stata program editor. When the program editor encounters the name of this program in its own line of code, the program editor will execute the commands in the program. There is one such program in the RD project to estimate proportions, called *proportion,* which is used repeatedly in the *dofiles*.

1.4.2 Dofile

A more complicated task, than one for a user-defined program, can be written into a dedicated *dofile* and called as required. One example employed in the RD project for analysis is RANDOM EFFECT MODELLING.do which executes the commands to construct the control limits of the funnel

plots (see below). Another is the BARCHART.do *dofile* that generates all the multiple bar charts in the final report. In executing lines of code, when the program editor encounters "do" followed by the name of the *dofile* (e.g., do BARCHART), it will execute all the **command**s in this *dofile*. If the *dofile* is not in the default folder, then the *dofile* should be preceded by its folder address.

1.4.3 Loops and macros

Another device for reducing repetitive code is the 'loop', which is used extensively in the RD project to perform the same analyses across different sub-groups of the cancer patient population defined by cancer site, demographic and clinical factors. A loop is a code that repeatedly executes the same command(s) on a list of elements defined in the header of the loop. In Stata loops are initiated with a *foreach* command that repeatedly sets a *local* macro name (see below) to each element of the list and executes the commands enclosed in braces. A macro is a named entity that stores some content (e.g. variable name) that is inserted into code when the name of macro appears accompanied by special punctuation. Stata has two types of macro: local and global. Local macros, for instance *x,* only work within *dofiles* and are recognised by the following punctuation: `x'. Global macros can operate across *dofiles* and are recognised by the following punctuation: ${x}. Loops set a local macro, e.g. *site*, include a list of elements (e.g. cancer site specific datasets), and a single set of commands. The command will have `site' embedded in the command where each of the elements' content will be inserted in turn and the command line executed. The set of commands is therefore written once only and available for review.

Macros are often used to allow flexible settings in a command. For instance, the y-axis label of a barchart may need to be set for either *Proportion (%)* or *Net survival (%)*. Macros, both locals and globals, are also used to allow flexibility in loops, where, for instance, a certain extra or different command is required for a certain sub-group of the patients (e.g. age group breakdown in survival analysis is defined differently for prostate cancer patients). This is accomplished by a conditional clause within the loop that, *if* the macro content equals the content in the element of interest in the loop, then the commands within the clause are performed.

2. Data preparation

A dedicated *dofile* called COMPILE DATASET.do produces the final or curated datasets for the RD project for the analyst to begin analysing. This means that the final clean variables to be analysed are defined (expressed in a variable label), as are their values (expressed in the value labels). COMPILE DATASET.do *dofile* is fully executable (i.e. can be run from start to finish) starting with calling up the base HBS dataset, and finishing with a set of curated data files representing the defined cancer site groupings.

COMPILE DATASET.do *dofile* also has code to import the English route-to-diagnosis estimates used to compare with NI. These have been extracted from spreadsheets publicly available on the PHE website; it is necessary to access or 'unhide' some of the Excel sheets where the base data that populates their interactive spread sheets lies. It is important to stress that no curation of the data set takes place outside of COMPILE DATASET.do *dofile*; this discipline ensures that data curation decisions are made and documented in the same place for easy review. COMPILE DATASET *dofile* has a set of comments, to assist in the data curation process.

3. Analytics

The main analytics employed in the RD project are the proportions of patients that are diagnosed through different routes-to-diagnosis, and the survival outcomes of patients within the different routes to diagnosis. Further analysis is required in the construction of the funnel plot control limits (Section 3.2).

3.1 Proportions

The proportion of patients going through different routes to diagnosis is calculated as the number diagnosed in that route divided by the total number of patients [though all routes]. The confidence interval is calculated which requires the standard error, SE(P) to be calculated, which equals √(P*(1-P)/n), where P is the sample proportion of a route-to-diagnosis, and n is the number of patients in all route-to-diagnosis of the sample proportion, P. The 95% confidence interval is calculated by the following formula (as follows) where $Z_{0.05}$ is the inverse of the cumulative density function set at 95%, and assuming P has a normal distribution: $P\pm(Z_{0.05}xSE(P))$.

Confidence intervals are used in the multiple bar charts as an indication of the precision of the estimates. Confidence intervals are indicative of where the true proportion, π, lies, therefore two non-overlapping confidence intervals are indicative of a significant difference between two proportions. A proper hypothesis test of the difference to be calculated in two proportions would require a specific test statistic about this difference and its standard error, accounting for the size of the proportions and the sample sizes. Moreover, if *a posteriori* multiple comparisons between proportions were to be undertaken a correction (e.g. Bonferroni) to the p-value of the test statistic would be needed to ensure a correct Type 1 error rate.

In order to have more meaningful comparisons between health geographies, the proportion of patients diagnosed via RDs by Trusts were age, sex and deprivation standardised. This was achieved by fitting a logistic model of an indicator (1=yes; 0=no) dependent variable for each RD against independent variables trust, age, sex and deprivation, and including interaction terms for trust by age. The overall or marginal proportion of for each Trust is then estimated by setting the Trust variable in the entire dataset to that particular Trust. Then for each individual Trust, a predicted proportion from the model (described above) is generated and these predicted probabilities are averaged over the entire dataset. In this way, each Trust's marginal proportion is standardised to the NI population.

3.2 Random effects model in proportions

A random-effects model was employed to construct the control limits of the funnel plots using a methodology described by Spiegelhalter (1). The modelling approach recognises two sources of variation, 1) random effects variation (τ^2) or variation between estimates about the population mean, μ, and caused by differing unmeasured case-mix factors, 2), sampling variation, s^2, which is largely a function of an estimate's sample size. The random effects variation, τ^2, and the pooled or weighted-proportion [of all regions], μ, are parameters that need to be estimated for the construction of the control limits; this estimation is carried out by the user-defined *metareg* (2) command in Stata.

Within the context of a set of [regional] proportions, the standard error (SE) for a particular estimate will be the square root of the sum of τ^2 and its sampling variation, i.e. $SE=\sqrt{(\tau^2+ s^2)}$; the formula for

the 95% and 99.8% control limits (without transformation of the data, see below) are $\mu \pm Z_{0.05}xSE$, and $\mu \pm Z_{0.001}xSE$, with $Z_{0.05}$ as described above. The funnel shape over the precision range is due to the changing ratio of s^2 to τ^2 in SE as s^2 decreases with increasing precision.

The random effects analysis was performed on the complementary log-log scale (3) meaning that all estimates and their sampling variation were transformed, the latter using the Delta method (3). The Delta method is a statistical technique to estimate the resulting variation of a random variable, x, after it has been transformed, e.g. log(x). The resulting standard errors estimated on cloglog scale for each estimate are then back-transformed to original scale resulting in control limits which lie between the theoretical bounds of a proportion, i.e. 0 and 1.

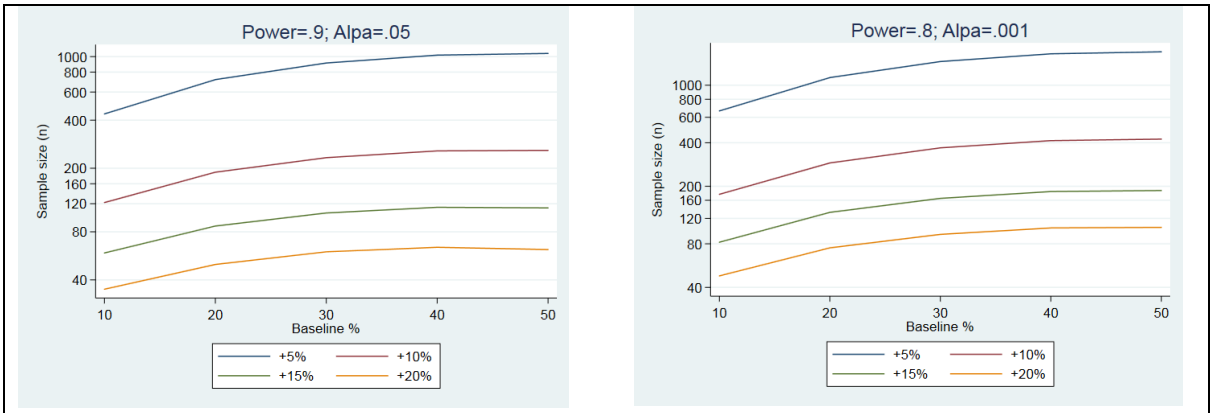
If there are outliers in the estimates, they inflate τ^2. Winsorisation (1) is a practice whereby the extreme estimates in the distribution of estimates are 'shrunk' towards the mean. Winsorisation at 5% was used to make parameter estimation more robust to outlying estimates, which were, through winsorisation, replaced by less extreme 95% percentile values of the estimate distribution about the pooled proportion.

3.3 Funnel plot at GP practice level

Figure 1A shows the sample sizes required in order to detect if a proportion in a GP practice departs by a certain amount from a baseline proportion, with Type I and Type II error rates set at 5% and 10%, respectively. For example, if the proportion of patients diagnosed through an emergency route-to-diagnosis was 20% in NI, then taking a particular GP-practice, it would require a sample approaching 200 patients in order to detect a proportion at 30% and reject the null hypothesis (no difference from 20%) ninety percent of the time.

The smaller the expected difference, the greater the sample size required to detect it (5% is the blue line in Figure 1A). Also, the sample size will increase as the baseline proportion approaches 50%. Due to the multiple comparisons implied in a funnel plot (GP practices, n= 345) the true Type 1 error rate is probably much higher that the set (5%) level. If the Type 1 error rate is decreased from 5% to 0.1% (equivalent to the 99.8% control limit), to protect against a false Type 1 error, and the Type 2 error rate is relaxed to 20% (this will offset the increase in sample size due to lower Type 1 rate), the sample sizes still increases considerably (Figure1B). This analysis assumed that there is no random effects variation, which would increase the variation in the data, above the sampling variation only considered here.

The total number of cancer cases in 2012-2016 was 46,000 which means an expected number per GP practice (n=345) is 133. This number is not sufficiently high to be able to detect realistic differences under +10%, and many GP practices' number would fall below this mean average. Maintaining a minimum and reasonable Type 1 and Type 2 error rates for all GP practices in a funnel plot is desirable from an equity point of view. In conclusion, the numbers of cancer patients per GP practice in the period 2012-2016 is too small to conduct a credible funnel plot analysis, and will require a larger diagnosis period to measure the true random effects variation between all GP practices in NI. Higher level geographies in NI, such as Integrated Care Partnerships (n=19), Local Government Councils (n=11), and Trusts (n=5) will have enough numbers in the main four cancer sites (5-6,000 cases) for the 300+ cases in these geographies.



A B

Figure 1: Sample size (n) required to detect proportions with a significant difference from a baseline proportion, with A) Type I (alpa) and Type II (I-power) error rates of 5% and 10%, respectively, with B) Type I and Type II error rates of 0.1% and 20%, respectively.

3.4 Survival

Three year net survival used the *complete* (4) approach to estimating the survival of cancer patients diagnosed from 2012-2016 and followed up to the end of 2017. The complete approach estimated survival at 3 years post-diagnosis, but without complete 3-year follow-up for each patient; more recently diagnosed patients (2015-2016) will only provide information up to 2-years of follow-up. The rationale for the complete approach is improved recency, i.e. more recent information on the survival experience in the early years of follow-up, than the *cohort* approach which require all patients to be followed up for the full 3-years. The complete approach still had at least 3 years (2012-2014) worth of patient data with a minimum 3-years of follow-up.

In future updates to RDs, it will be possible to extend follow-up and produce, as standard, five-year estimates. The net survival estimator used was the Pohar-Perme (5), implemented in a user-led program, *stns* (6) by the Stata statistical software, Version 14. Net survival is a theoretical statistic of the survival of patients if all non-cancer background mortality was removed. Net survival of cancer patients is driven only by the excess mortality rates of the cancer patients, and allows comparison between groups whose background mortality rates may differ. The Pohar-Perme estimator is a non-parametric estimator, like the Kaplan-Meier observed survival estimator, that estimates net survival by inverse-weighting each patient's contribution to death rates by their expected survival; this has the effect of removing their background or non-cancer mortality rates.

Mortality rates of NI lifetables are used to approximate the non-cancer mortality rates of the NI cancer patients as the proportion of deaths caused by a single type of cancer are proportionally very small to the total. To prevent imprecise survival estimates, at least 50 patients were required to contribute data; this approach reflects practice in international cancer studies such as CONCORD (7). Within specific cancer sites or groupings, only a patient's first tumour was used in survival analysis. Patients diagnosed via death certificates or post-mortem were excluded as their 'true' diagnosis date was not properly established or found.

4. Presentation of results

4.1 Tables

4.1.1 Proportions

The frequency (n) and proportions (%) of patients diagnosed within each route to diagnosis are presented in tables in the report (e.g. Table 1). The numerator of the proportion represents the number of patient diagnosed in a particular RD, while the denominator is the combined number of patients through *all* the routes including unknown and DCO RDs. The Honest Broker Service rule to protect against the disclosure of potentially-identifiable patient information is that no cell of a table should contain a frequency <10. In addition, where only one cell in a table had <10, another RD, usually DCO or unknown, was also removed from the table to prevent calculating the number in the one concealed cell by subtracting, from the total, the numbers in the unconcealed cells.

Table 1: The frequency (n) and distribution (%) of routes-to-diagnosis of breast cancer patients, by year of diagnosis

Route-to-diagnosis	Year of diagnosis									
	2012		2013		2014		2015		2016	
	n	%	n	%	n	%	n	%	n	%
Screening-RD	406	30.8	357	27.5	374	28.9	430	29.4	412	28.7
Redflag-RD	623	47.2	666	51.3	644	49.7	752	51.4	689	48.0
GP-RD	110	8.3	102	7.9	93	7.2	86	5.9	127	8.9
Outpatient-RD	84	6.4	75	5.8	82	6.3	98	6.7	105	7.3
Emergency-RD	52	3.9	63	4.8	51	3.9	52	3.6	64	4.5
Total	**1,319**		**1,299**		**1,295**		**1,463**		**1,434**	

Note: frequency (n) will not sum to total due to exclusion of DCO-RD and unknown-RD

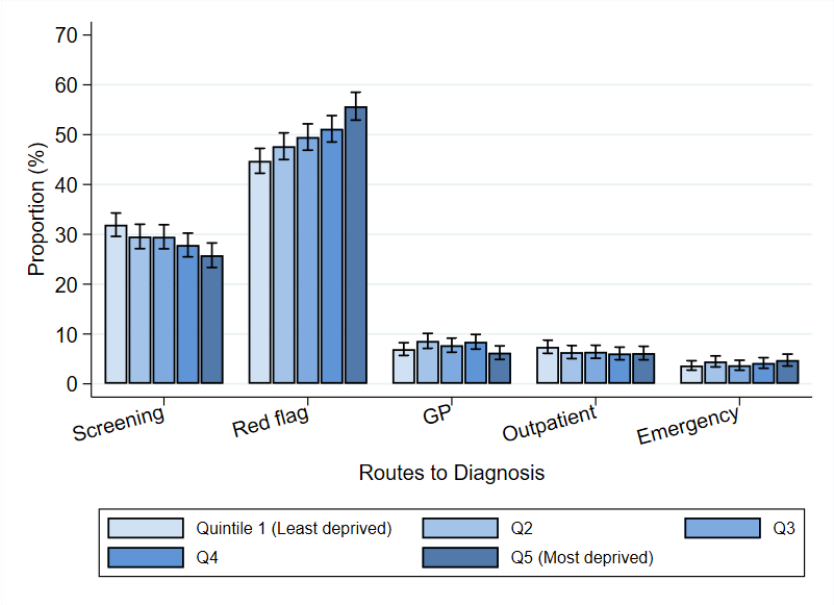
The tables are initially designed in the Microsoft Word report draft. In Stata, the frequency and proportions, expressed in percentages, are formatted (e.g. number of decimal places, commas in large integers), turned into strings, restructured to fit the table format in the Word draft, and exported to an excel file. The strings are prefixed with a tilda ("~") to preserve the leading zeros in Excel. After assessing the cell numbers, the RDs that need to be concealed in the table (e.g. RD with <10 patients) are deleted from the Excel file. The remaining block of data from the Excel file is copied into the Word table, the leading tildas removed by a 'replace' command, and the table format is applied using the Format Painter. Report tables are populated in this way in order to minimise disclosure and transcription errors.

4.2 Graphing

4.2.1 Multiple Bar Charts

The multiple bar chart is used to compare the distribution (%) of RD categories between different groups of the cancer patient population. Groups can be defined by demographic and clinical factors (stage of disease). For instance, the female breast cancer patient population can be divided into 5 quintiles of deprivation, and the proportion of screening RD declines with increasing deprivation (Figure 1). In the multiple bar chart, the RD is a random variable that has potentially eight categories, and the factors define groups (or by-groups) of patients in the population, e.g. deprivation quintile. The multiple barchart sets together, from the different by-groups, bars of the same category of the random variable, thus making comparisons, such as trends or gradients, easier. The interpretation of these comparisons is aided by the confidence intervals—an indication of precision—about the bars (see Section 3.1).

Figure 2: Distribution (%) of routes-to-diagnosis for patients diagnosed with breast cancer within Northern Ireland 2012-2016, by deprivation quintile



Stata does not have a simple multiple barchart command (with confidence intervals) required for the RD project. Instead, it was necessary to use the *twoway* graphing command in Stata to amalgamate several by-group specific barcharts into the one graph. To be able to generate multiple barcharts, an algorithm was required to meet the following requirements:

1) A set of numbers on the x-axis is assigned to each by-group to situate their [random variable] category bars (and confidence intervals), in such a manner that like-category bars, across the different by-groups, are grouped together in a consistent order (Figure 1). Also, a set of numbers are reserved to leave space between the groups of like-category bars on the x-axis. To cut down white space on the top of the chart, an upper range on the y-axis was set at 10% greater than the greatest observed proportion and the rounded to the nearest multiple of 10%.

2) The x-axis requires a label for each like-category group (e.g. Screening) and the label's position to be in the centre of the group on the x-axis.

3) The legend requires the correct assignment of random variable's category value names, e.g. Quintile 1 (Least deprived). Also, there is an option to suppress the legend when there is only one by-group.

The *dofile* BARCHART.do is designed to meet the requirements above with the flexibility that the number of categories (RDs) may change from site to site, and the number of by-groups may change. In some cases, for natural reasons, such as no screening RD in lung cancer, or because the category has been suppressed due to low numbers. In Figure 1 for female breast cancer, there are five by-groups for level of deprivation each with five random variable categories presented: screening, redflag, GP, outpatient and emergency. The x-axis has a space at multiples of 6 to separate off the like-categories from each by-group. The category labels are positioned in the middle of the like-category groups. The y-axis is restricted from 0 to 70%.

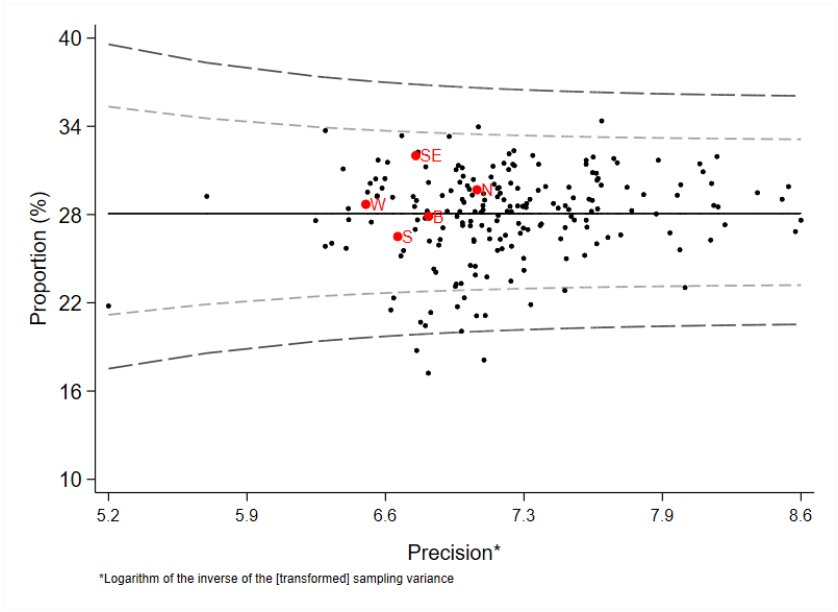
4.2.2 Funnel plot

Figure 2 shows a funnel plot that was used to assess whether the proportion of breast cancer patients diagnosed through screening in the NI Trusts was in keeping with observed variation in English Clinical Commissioning Groups (CCGs). The black solid line represents a population mean parameter, μ, from the random effects modelling (Figure 2). The short- and long-dashed lines in a funnel shape represent the 95.0% and 99.8% control limits about the population mean, respectively. The control limits are constructed using the estimates' standard error which is a combination of the estimate's sample variation and the random effects variation (τ^2) from the random effect model; the latter represents the CCG variation in the proportion screened that is not due to sample size.

The random effect modelling was completed on the complementary log-log (cloglog) scale (see above under analysis) which required, using the Delta method (3), the sampling variation to be transformed onto the cloglog scale. The cloglog-transformed sampling variation is not solely a function of the sampling variation, but also the estimate's proportion, and therefore there is no monotonic relationship between sampling and transformed-sampling variation. This means that in order to have smooth [back-transformed] control limits, which are functions involving the transformed standard errors, it is necessary to use the inverse of the *transformed* sampling variation as a measure of precision. Presenting this precision on the logarithm scale ensures a more even spread of the observed values on the graph.

In a process control setting, when an estimate breaches the 95% control limits it is a warning, whereas breaching the 99.8% control limit signifies a real departure from the population mean. Adopting this conservative approach protects against making Type 1 statistical errors.

Figure 3: Proportion of breast cancer patients diagnosed through a screening route-to-diagnosis for English Clinical Commissioning Groups (n=195, 2006-2016), and Northern Ireland Trusts (n=5, 2012-2016) in a funnel plot



5. Glossary

Bonferroni test: is a type of multiple comparison test used in statistical analysis. When performing a number of hypothesis tests with multiple comparisons, eventually, a result could occur that shows statistical significance of the dependent variable, even if there is none.

Code: instructions or commands in computer program.

Command: an instruction causing a computer to perform one of its basic functions.

Dependent variable: the outcome variable that is related to a set of independent variables in a regression model. The dependent variable will be to the left of the equal sign in the model formula.

Estimate: an approximate calculation of the value, number, quantity, or extent of something.

Estimator: a method for arriving at an estimate of the value of a parameter.

Folder (default): a place where statistical package will first look for data files or by default save any generated files.

Folder: a place in a computer where files are stored.

Funnel plots: are recommended as a graphical aid for regional comparisons, in which regional estimates of an underlying quantity are plotted against an interpretable measure of its precision. 'Control limits' form a funnel around the target outcome or overall mean, estimates that breach control limits are considered statistically significantly different from the overall mean.

Hypothesis test: Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. Hypothesis testing is used to infer the result of a hypothesis performed on sample data from a larger population.

Independent variables: set of variables that is related to an outcome or dependent variable, usually in the context of a regression model. Independent variables will be to the right of the equal sign of the model formula.

Kaplan-Meier estimator: also known as the product limit estimator, it is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment.

Life Table: In actuarial science and demography, a life table (also called a mortality table or actuarial table) is a table which shows, for each age, what the probability is that a person of that age will die before his or her next birthday ("probability of death").

Logarithm: a quantity representing the power to which a fixed number (the base) must be raised to produce a given number.

Monotonic: a function varying in such a way that it either never decreases or never increases.

Multiple bar chart: (also known as grouped or clustered **bar charts**) are used to present and compare data of sub-categories within the main category. In other words, unlike singe series **charts**, in this **chart** type, each category has two or more than two data series.

Net survival: the proportion of patients alive at a certain time point after treatment when only considering their deaths arising from their condition, i.e. excess deaths.

Observed survival: the proportion of patients alive at some time point after their diagnosis.

Program editor: part or function of Stata that executes submitted program syntax, and provides results and feedback to a results window.

Program: a series of coded instructions to control the operations of a computer.

Quintile: any of five equal groups into which a population can be divided according to the distribution of values of a particular variable.

Random variable: A random variable is a variable or a function that assigns values to each of an event's outcomes.

Scripted syntax: A file written by an analyst with designed structure of code in a computer language.

Standard operating procedure (SOP): is a set of step-by-step instructions compiled by an organization to help workers carry out complex routine *operations*. SOPs aim to achieve efficiency, quality output and uniformity of performance, while reducing miscommunication and failure to comply with industry regulations.

Standardisation by age, sex and deprivation: a method where by a statistic, e.g. a proportion, is adjusted to a standard population of the same age, sex and deprivation structure.

String: is a contiguous sequence of symbols or letters, such as a character string (a sequence of characters .

Type 1 error: the probability of rejecting the null hypothesis when it is true.

User-defined: refers in this SOP to a program that is run on Stata but was not written by Stata but by a user of Stata.

6. References

1. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Stat Med. 2005 Apr 30;24(8):1185–202.

2. Harbord RM, Higgins JP. Meta-regression in Stata. Meta. 2008;8(4):493–519.

3. Quaresma M, Coleman MP, Rachet B. Funnel plots for population-based cancer survival: principles, methods and applications. Stat Med. 2014 Mar 15;33(6):1070–80.

4. Brenner H, Gefeller O. Deriving more up-to-date estimates of long-term patient survival. J Clin Epidemiol. 1997 Feb;50(2):211–6.

5. Perme MP, Stare J, Esteve J. On estimation in relative survival. Biometrics. 2012 Mar;68(1):113–20.

6. Clerc-Urmés I, Grzebyk M, Hédelin G. Net survival estimation with stns. Stata J [Internet]. 2014;14(1):87–102. Available from: http://www.stata-journal.com/article.html?article=st0326

7. Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang X-S, et al. Global surveillance of cancer survival 1995-2009: Analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). Lancet. 2015;385(9972).