The THAI Journal of SURGERY 2021;42(4):161-166.
Official Publication of The Royal College of Surgeons of Thailand

Original Article

SURPY Python Toolkit for Data Analysis

Surasak Sangkhathat, MD, PhD[*,†]
Wison Laochareonsuk, MD[†,‡]
Komwit Surachat, MSc, PhD[§]

[*]*Division of Surgery, Faculty of Medicine, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand*
[†]*Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand*
[‡]*Division of Biomedical Science and Biomedical Engineering, Faculty of Medicine, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand*
[§]*Division of Computational Science, Faculty of Science, Prince of Songkla University, Hatyai, Songkhla 90110, Thailand*

Abstract ***Objective:*** SURPY is a Python-based package for statistical analysis available on PyPi repository. The present study aims to evaluate performance of the SURPY package in providing basic data analysis compared to a standard statistical package, Stata v.14 (StataCorp, College Station, TX, USA).

Methods: Datasets from previously published studies were retrieved for analysis. The data was transferred to the .DTA format for analysis using the Stata v.14 program and was imported as a dataframe into the Python 3.0 environment, to be analysed by the 'soap' (surgical outcome analysis program) package of SURPY 1.1.7. Results of the analysis from the 2 programs were compared.

Results: The soap package from the SURPY program was able to import data stored in the Microsoft Excel format and calculate basic descriptive statistics. The program correctly performed *t*-tests and Mann-Whitney U tests. Also, the program was able to produce Kaplan-Meier survival curves and perform log-rank tests, which gave similar outputs compared to those from the Stata program.

Conclusion: The SURPY program can be used for simple data analysis, which could be useful for surgeons who are not familiar with typing commands in commonly used statistical programs. The SURPY program can be further developed to incorporate graphic user interface.

Keywords: Data analysis, Python program

Introduction

Various computer packages are available for surgical data analysis, such as Stata, the R program, and Microsoft Excel. However, most of these programs require a certain level of computing skills and might not be friendly enough for the average surgeon to use. Also, some statistical packages are not suitable for datasets with large dimensions.[1]

Python is a high-level programming language created in the late 1980s by Guido van Rossum and his colleagues in the Netherlands.[2] As the program is developed under an open-source license, and is currently administered by the Python Software Foundation, it is free to use, modify and distribute. The python interpreter, now version 3.9.0 (October 2020), is free to download into most operating systems, including

Received for publication 27 February 2021; Revised 16 July 2021; Accepted 18 July 2021

Correspondence address: *Surasak Sangkhathat, MD, PhD, Division of Surgery, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla, Thailand 90110; Email: surasak.sa@psu.ac.th*

Windows, Linux and Mac OS.2 The repository hub for Python open-source packages is PyPI (https://pypi.org). Python has become more popular in recent years among data scientists, especially for statistical analysis, natural language processing and machine learning.[3,4] In the health care services, Python packages have been built for various purposes, such as medical image analysis, machine learning for risk prediction, clinical trial data management and bioinformatics.[5-8]

SURPY is a Python-based program, developed by the authors, designed for surgical data analysis and ease of use. The program includes packages for data description, statistical testing, tests for association, and simple survival analysis. Surgical data analysis usually includes data summary, analysis for association between variables and data displays with graphs and tables. The present study aims to compare the results of simple data analysis between SURPY and Stata programs.

Materials and Methods

SURPY began as a Python project to provide a simple tool for basic statistical analysis in surgery. The first version was launched in PyPI on January 25, 2021. The latest version, SURPY 1.1.7 released on April 8, 2021, was used in the present study. The 'soap' package in the SURPY at that time included sub-programs for data description (soaplore(data, outcome), parametric and non-parametric tests between 2 groups (soap_TU(data,outcome,variable)), soap_multi_T(data, outcome) [tests of multiple quantitative variables between groups defined by a fixed binary variable] and soap_T_for_multibinary(data, variable) [tests of a fixed quantitative variable between groups defined by multiple binary variables], chi-square tests of association between 2 categorical variables (soap.soap_x_tab(data, variable_a, variable_b)) and (soap_x_across(data,outcome)), and Kaplan-Meier survival curve estimation (single_kmc(data, status, interval)) as well as comparisons between survival curves (compare_kmc(data, factor, status, interval)) (Figure 1).

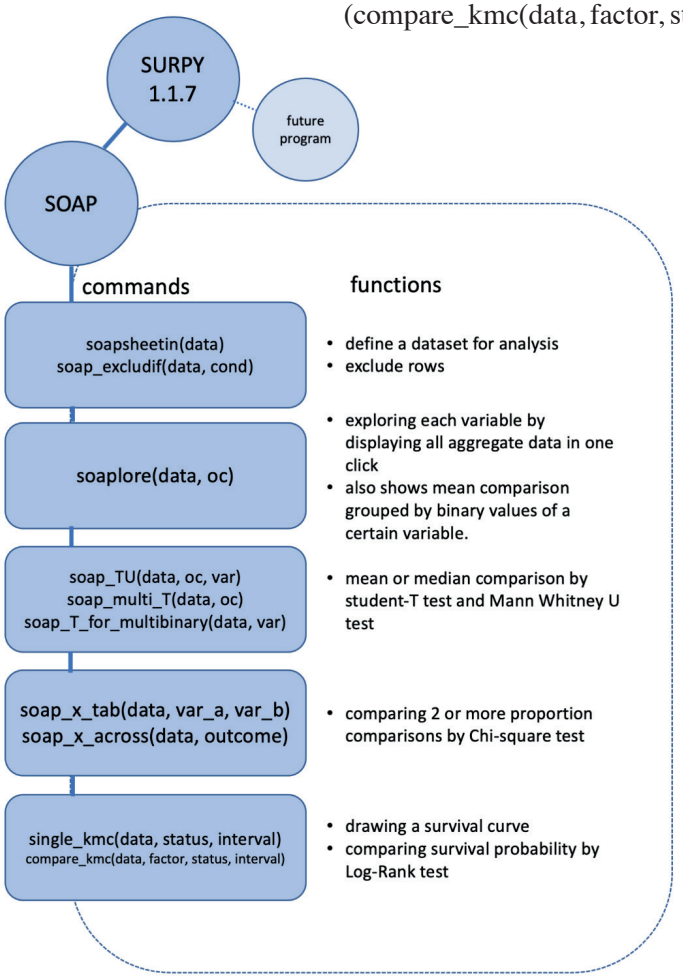


Figure 1 Components of the soap program in SURPY package version 1.1.7 released April 8, 2021

Datasets derived from previously published studies by the authors were used in the present study for re-analysis using both the SURPY program version 1.1.6 and a reference statistical program, Stata version 14 (StataCorp, College Station, TX, USA). The results of this analysis were compared in terms descriptive or summary statistics, p-values from t-test and chi square tests, and Kaplan-Meier plots.

Results

As all datasets were originally stored in the Microsoft Excel format, they were converted to the DTA file format (filename.dta) by using the StatTransfer version 12 program. Accessing the SURPY package can be done through downloading from PyPi (https://pypi.org/project/SURPY/) (Figure 2) and the manual can be found on https://github.com/sasurasa/Surgical-Outcome-Analysis-on-Python/blob/SURPY/SURPY%20manual%20190321SS.pdf. SURPY correctly displays summary statistics for each variable in the dataset. The program can also compare mean values of quantitative variables between categories of binary variables (Figure 3).

In significance testing of quantitative variables between 2 groups, either Student's t test or Mann-Whitney U test are commonly used. The soap package on SURPY can perform these tests well. All p-values obtained using the package were similar to those calculated by the reference program. In addition, the package can display Box-and-Whisker plots for visual comparison between groups (Figure 4).

When a researcher needs to perform multiple comparisons of quantitative data between categories of a fixed binary variable, the code 'sp.soap_multi_T(data, oc)' returns the same results as when using the Stata program. Similarly, performing comparisons of a single quantitative variable between categories of multiple binary variables could be done with one click using the command 'sp.soap_T_for_multibinary'.

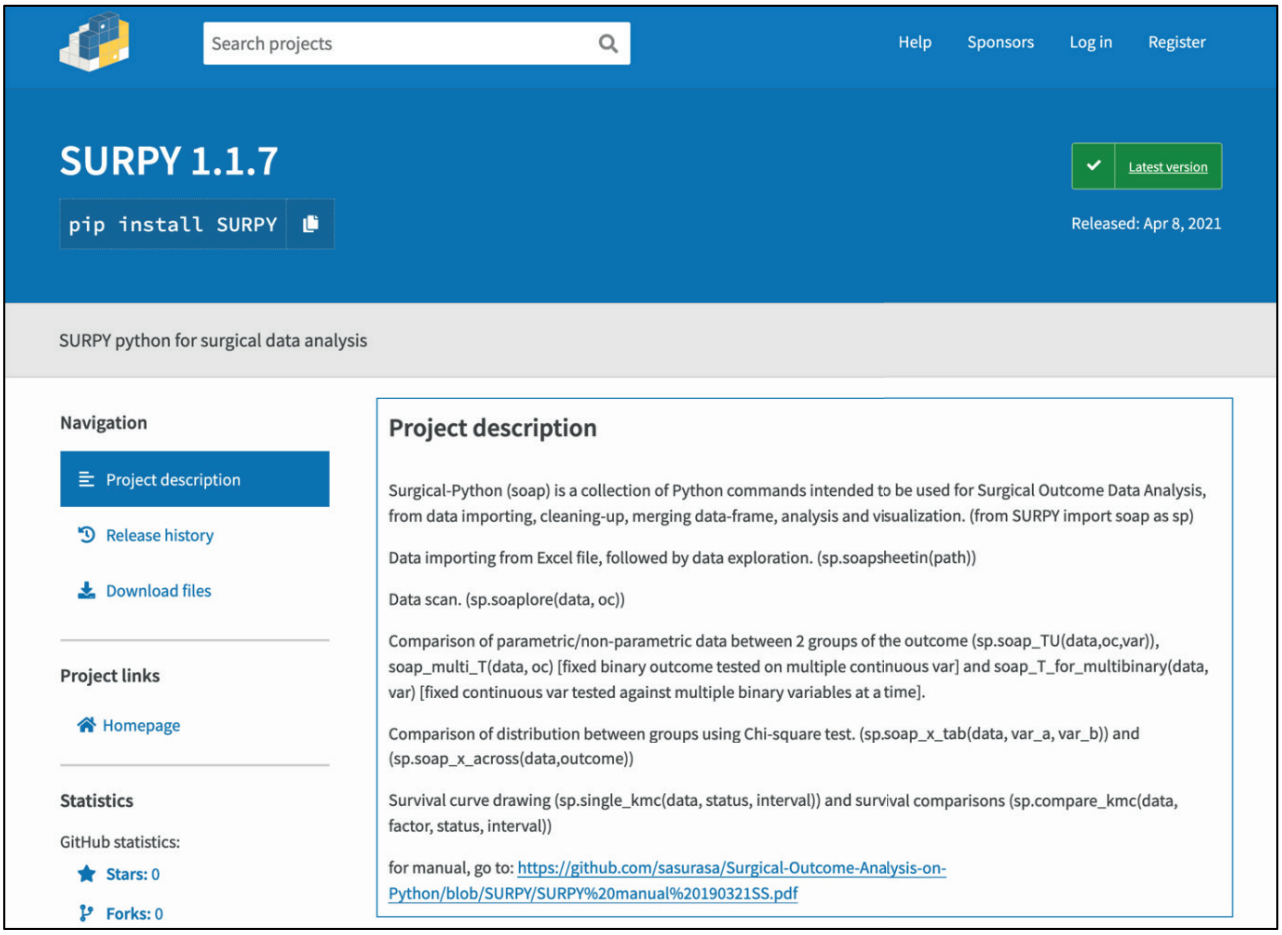


Figure 2 The repository page of the SURPY program which can be accessed at https://pypi.org/project/SURPY/

A

	bh	cea	modesx	operation	treatyear
count	716.000000	1112.000000	1159.000000	1162.000000	1164.000000
mean	196.500000	163.750899	0.073339	5.067126	2008.759450
std	373.926905	728.424751	0.279968	3.946320	2.585535
min	113.000000	0.000000	0.000000	0.000000	2004.000000
25%	155.000000	4.000000	0.000000	3.000000	2007.000000
50%	161.000000	17.000000	0.000000	4.000000	2009.000000
75%	168.000000	88.000000	0.000000	7.000000	2011.000000
max	7174.000000	10500.000000	4.000000	15.000000	2014.000000

	surgeon	ajcc	lhs	fstchm	totcyc
count	1160.000000	1158.000000	1149.000000	1135.000000	797.000000
mean	17.174138	2.702936	14.280244	4.103084	5.336261
std	26.437626	0.884814	8.777732	2.647574	4.627029
min	0.000000	1.000000	2.000000	0.000000	0.000000
25%	8.000000	2.000000	10.000000	1.000000	2.000000
50%	8.000000	3.000000	12.000000	4.000000	6.000000
75%	12.000000	3.000000	16.000000	7.000000	8.000000
max	99.000000	4.000000	88.000000	8.000000	36.000000

B

sex	cea	modesx	operation	treatyear	surgeon	ajcc
0	185.031509	0.070288	5.022364	2008.771930	16.982400	2.664526
1	138.540275	0.076923	5.119403	2008.744879	17.398131	2.747664

sex	lhs	fstchm	totcyc	targetx	ocmtcour	opyear
0	14.672609	4.117550	5.108747	25076.190231	7.543750	2008.771930
1	13.825188	4.086629	5.593583	20683.046703	7.491018	2008.744879

Figure 3 Output of the command sp.soaplore(data, 'sex') displaying descriptive statistics; (A) summary of each variable displaying the mean, median, minimum, maximum, standard deviation and interquartile range; (B) displaying mean values of quantitative variables according to categories of a binary variable (sex in this case) for comparative purposes.

```
>>> sp.soap_TU(data, 'ev', 'lsm')
lsm
     count       mean        std   min    25%   50%   75%   max
ev
0    15.0  17.366667  20.646053   3.0   5.55   8.2  14.5  75.0
1     5.0  33.700000  25.119415  11.3  16.80  28.9  36.5  75.0
------------------------------------------------------------

Ttest_indResult(statistic=-1.4562389048044695, pvalue=0.16254863419294766)
MannwhitneyuResult(statistic=14.5, pvalue=0.024724498879759953)
```

A

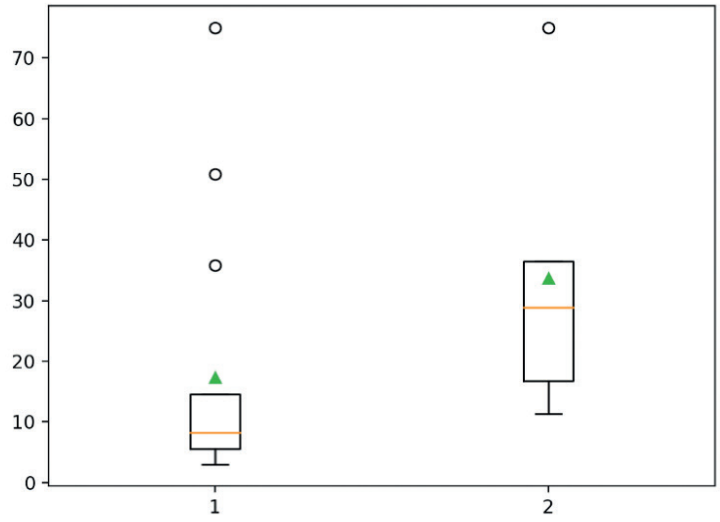


Figure 4 Demonstrating (A) the comparison of the means and medians between 2 groups, as well as performing the *t*-test and Mann Whitney U test; (B) Box-and-Whisker plot of a comparison of a quantitative variable between 2 categories, which shows values of the median (orange lines), the mean (green triangles), interquartile range (box), whiskers, and outliers (circles)

```
>>> sp.soap_x_tab(data, 'ev', 'portalht')
================================================================================

Data dimension:  (2, 2)
Chi-square value =  0.0
Chisquare p-value = 1.0
================================================================================



         ev  portalht0  portalht1  horizonsum  %portalht0  %portalht1
0         0         14          1          15   93.333333    6.666667
1         1          4          1           5   80.000000   20.000000
2  vertisum         18          2          20   90.000000   10.000000
```

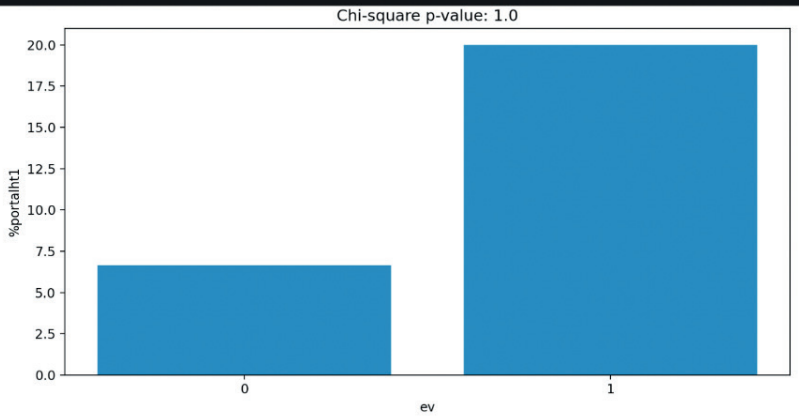
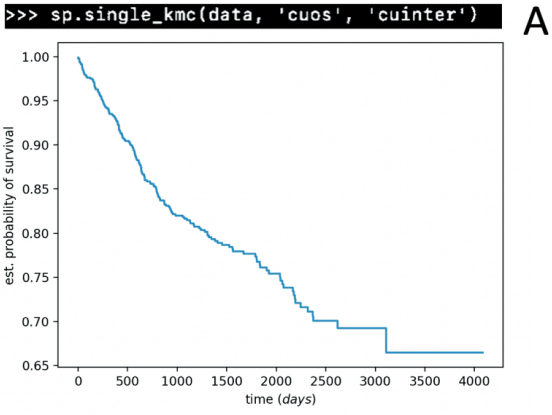
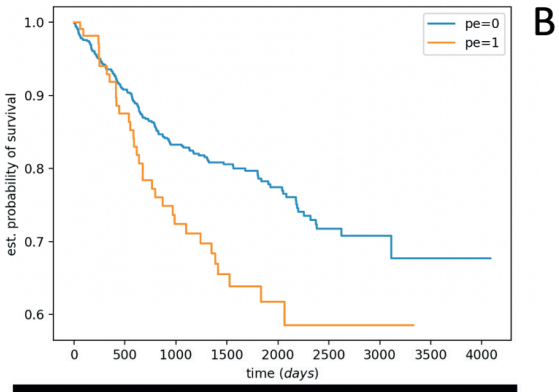


Figure 5 Demonstrating a 2 x 2 table, results of chi-square testing and p-values, and a bar-chart showing percentages






```
test_statistic    p  -log2(p)
          7.70 0.01      7.50
```

Figure 6 Showing Kaplan Meier survival curves produced by the soap program on SURPY 1.1.7 for: (A) a single group; (B) for 2 groups with a log-rank test statistic.

Performing cross tabulation of any 2 binary categorical variables or chi-square tests for 2 binary variables using the code 'sp.soap_x_tab(data, var_a, var_b) gave the same p-values as those derived from Stata (Figure 5). In addition, the soap program correctly calculated chi-square tests for all pairs of categorical variables in the dataset.

Similar to Stata's sts command, the SURPY program can display simple Kaplan-Meier survival curves (Figure 6). In addition, the program also correctly returns p-values of the log-rank test. However, SURPY has not been scripted for comparing differences in survival probabilities for 3 or more categories. In addition, it has not been scripted for customization of the graphs.

Discussion

The SURPY program is an initiative by the authors to develop a clinician-friendly platform for statistical analysis of large datasets. These first versions of the package were to help clinicians calculate basic summary statistics, perform simple statistical or significance tests, and to do basic survival analysis with Kaplan-Meier estimates and log-rank tests These requirements were derived from our experience in teaching data analysis for surgical trainees.

Python is an open-source interpretation language that has gained increasing popularity for use in writing programs for analyzing large, complex data sets.9 In the health sciences, python is employed as a programming language for various data-driven uses including large longitudinal data mining, deep learning, predictive modelling and high-throughput genomic data analysis.[10-12]

The SURPY package is aimed towards basic statistical analysis. Although typical clinical or surgical datasets can be managed effectively with various available statistical packages, we found that many surgeons avoided performing data analysis themselves because of the perceived difficulty in using these available packages. The goal of SURPY is to facilitate basic data analysis once a dataset is uploaded and types of variables in the data are defined. With a couple of clicks, all statistical associations of interest within the datasheet will be displayed in forms of networking diagrams and graphs.

The 'soap' is our pilot release of a more comprehensive toolkit, to demonstrate proof of concept. The program returns basic summary statistics across the sheet, as well as all statistical associations when categorical outcomes of interest are defined. One limitation of the current program is that the script does not automatically discriminate between quantitative or categorical numerical values. For tests of association, the program is scripted to identify a numerical variable with 5 or less unique values as a categorical variable.

To provide a comprehensive solution for all surgical data analysis, the program requires continuing development. Notably, functions for multivariate analysis, customized tests of association and a graphic user interface need to be provided to enhance the usefulness of the program.

Conclusion

In summary, the soap package of the SURPY program was developed for surgical data analysis. The release of the program and benchmarking against a standard statistics program have proven its original concept that automation of data processing is feasible.

Acknowledgement

Dave Patterson from the Division of Foreign Affairs edited the English language in the manuscript.

References

1. Grigis A, Goyard D, Cherbonnier R, et al. Neuroimaging, genetics, and clinical data sharing in Python using the CubicWeb Framework. Front Neuroinform 2017;11:18. doi:10.3389/fninf.2017.00018.
2. Gowrishankar S, Veena A. Introduction to Python programming. Bota Raton, FL: CRC Press, Taylor and Francis Group; 2019.
3. Lee GH, Shin SY. Federated learning on clinical benchmark data: performance assessment. J Med Internet Res 2020;22:e20891. doi:10.2196/20891.
4. Raschka S, Kaufman B. Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition. Methods 2020;180:89-110.
5. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. Value Health 2019;22:808-15.
6. Semeraro R, Magi A. PyPore: a python toolbox for nanopore sequencing data handling. Bioinformatics 2019;35:4445-7.
7. Albers PN, Wright CY. Clinical trial data management in environmental health tailored for an African setting. Int J Environ Res Public Health 2020;17:doi:10.3390/ijerph17020402.
8. Jungo A, Scheidegger O, Reyes M, et al. A Python package for data handling and evaluation in deep learning-based medical image analysis. Comput Methods Programs Biomed 2021;198:105796. doi:10.116/j.cmpb.2020.105796.
9. Niewinski G, Smyk W, Graczynska A, et al. Kidney function after liver transplantation in a single center. Ann Transplant 2021;26:e926928-1-8.doi:10.12659/AOT.926928.
10. Kuntzelman KM, Williams JM, Lim PC, et al. Deep-learning-based multivariate pattern analysis (dMVPA): a tutorial and a toolbox. Front Hum Neurosci 2021;15:638052. doi:10.3389/fnhum.2021.638052.eCollection 2021.
11. Liu G, Lu D, Lu J. Pharm-AutoML: an open-source, end-to-end automated machine learning package for clinical outcome prediction. CPT Pharmacometrics Syst Pharmacol 2021;10:478-8.
12. Wu Z, Wang X, Pan R, et al Study of the relationship between ICU patient recovery and TCM treatment in acute phase: a retrospective study based on Python data mining technology. Evid Based Complement Alternat Med 2021;2021:5548157. doi:10.1155/2021/5548157.