University of Louisville
Journal of Respiratory Infections

ORIGINAL RESEARCH

Distributing Data and Analysis Software Containers For Better Data Sharing in Clinical Research

*William A Mattingly[1], Stephen P Furmanek[1], Christopher Sinclair[3], and Timothy L Wiemken[2]

Abstract

Introduction: Data sharing in clinical research is critical for increasing knowledge discovery. Data and software tools should be FAIR: Findable, Accessible, Inter-operable and Re-usable. Many bottlenecks exist in the process of a clinical investigator using shared data including data acquisition and statistical analysis. The objective of this project is to develop a structure for sharing data and providing rapid automated statistical analysis through creation of a pre-packaged, open-source software container.

Methods: We use the open source software container technologies VirtualBox and Vagrant to create a template for sharing clinical data and analysis scripts as a single container. We use a timer to record the time necessary to setup and initialize the software container and view the results.

Results: We have created a template for sharing data and analysis scripts together using open source software container technologies VirtualBox and Vagrant. We found the time needed to initialize the container to be 5 minutes and 36 seconds for a macOS-based machine and 7 minutes and 2 seconds for a Windows-based machine. Containers can be downloaded and executed from any Mac or Windows computer allowing both the reuse of and interaction with the data. This greatly reduces the time and effort needed to obtain and analyze clinical data.

Conclusion: Reducing the time and effort needed to obtain and analyze clinical data increases the time available for data exploration and the discovery of new knowledge. This can be effectively achieved using software containers and virtualization.

DOI: 10.18297/jri/vol1/iss4/6
Received Date: July 31, 2017
Accepted Date: September 13, 2017
Website: https://ir.library.louisville.edu/jri
Affiliations:
[1]University of Louisville
School of Medicine
Division of Infectious Diseases
[2]University of Louisville
School of Public Health and Information Sciences
Department of Epidemiology and Population Health
[3]University of Louisville
School of Business
Department of Computer Information Systems

Introduction

For many years, there has been a growing need for data management standards for the sharing and reuse of research data. Public data sharing policies have been a part of government funded research for many years [1], and several organizations have recently reiterated this importance as technologies continue to make data more accessible [2-5]. Data collected and generated by investigators is often stored in an ad-hoc fashion, with a structure that is clear and consistent to the investigator and research team, but not necessarily by those who may be interested in its reuse. This is especially important to public and private funding organizations, where data are the product of an investment and must continue to have value into the future. “Data stewardship” is a common term used to describe this new trend for researchers structuring their data to support future use.

Recently, the NIH and other public funding bodies have adopted the FAIR principles [6] as a general guideline for the necessary features needed to facilitate data sharing. These features include Findability, Accessibility, Interoperability, and Reusability. In this paradigm, not only is it important that data be structured for reuse by other investigators, but also structured for machine and software interfaces as well. More and more data are being accessed by software data mining and discovery platforms, and each requires consistent and standardized data structures to be effective at knowledge discovery. Fortunately, data structures designed to be machine-readable can be enhanced to support human readability as well. The development and adoption of these new standards will be a recurring theme in the future of research.

In addition to making raw research data accessible, FAIR principles are intended to apply to the software that researchers use to analyze their datasets. This has led to the concepts of data authorship and research objects. [7, 8] Research objects can include the analysis software code used to generate results in addition to the dataset itself. Creating these structures can be challenging in terms of time spent by investigators [9]. It is also cumbersome to make shared software analysis code reusable. The efficient reuse of software source code is a focus of the discipline of software engineering [10], and effort must be invested by programmers early in the development process for software to be reusable. Without this effort, it takes more time to understand the intent of the original programmer than to write a new program. Modern programming languages have made it easier to apply the principles of software reuse and even novice programmers can now develop software that is easy to extend,

*Correspondence To: William A Mattingly, PhD
501 E Broadway, Suite 120B
Louisville, KY 40202
bill.mattingly@louisville.edu

modify, and reuse [11]. In the area of clinical research, following FAIR principles continues to be a challenge. Furthermore, little work has been done to make it simple for clinical investigators to use these principles in obtaining and analyzing their data.

From informal interviews with pneumonia researchers and statisticians we found several obstacles to creating shared datasets in this field. Two obstacles stand out from the others. The first was the difficulty in giving the data the appropriate context to be interpreted accurately by subsequent investigators. This context can consist of the specific features of the study population, the conditions under which the data was collected, and the types of research questions the data was gathered to answer. The second major obstacle is the time and effort needed to replicate the analysis pipeline used by the primary investigator. These pipelines can be very sophisticated and their setup can be time consuming to replicate. If this setup could be automated it could improve the ability of shared data to be used by others.

Our objective was to improve the utility of shared datasets by creating a fast and easy to use software container for sharing research data and statistical analyses. This container will support the FAIR principles of findability, accessibility, interoperability, and reusability. We record the startup time needed for the software container and describe the steps necessary for its setup and execution. The container will also support the addition of contextual information about the data in the form of documentation and commentary from the creators.

Terms and Abbreviations	
OS - operating system	The software for managing interactive programs on a computer.
virtual machine	Software that partitions physical hardware into virtual hardware that can run contained software environments.
FAIR	Findable, Accessible, Interoperable, Reusable
data stewardship	The facilitation of data re-use by researchers and investigators
virtualization	the process of running software inside a virtual machine
VirtualBox	An open source software virtualization program
open source software	Software that is made freely available with little to no licensing restrictions
Vagrant	open source virtual machine management software
R	An open source programming language supporting many statistical tests
Linux	A popular open source operating system
proprietary software	software that has licensing restrictions governing its use and distribution

Methods

Data used in this study originate from the University of Louisville Pneumonia Study, a three-year study on the incidence, epidemiology, and clinical outcomes of hospitalized patients with community-acquired pneumonia[12]. This study took place from June 1, 2014 to March 31, 2017.

When designing the software container, we set out to address each of the four FAIR principles to the best of our ability. How an investigator addresses FAIR principles when sharing data will depend upon many factors, such as the type of data being shared and the type of software used to analyze data. For these reasons, the methods used for this study may not translate in their entirety to other studies. We describe below the FAIR principle and how it was addressed.

1. Findability: Data should be easy to find. For this study we used Zenodo[13], a free online service funded by CERN[14] to generate a DOI or permanent document object identifier, for our dataset and software container. Zenodo registers DOIs through DataCite, and provides means for updating and retracting incorrect data[15].
2. Accessibility: Data should be easy to access. Our data is de-identified and will be hosted online along with the software container. Any user with an internet connection can access it.
3. Interoperability: Data should be in a standardized format. We share our data in a comma separated value file with a header row describing the variable name. This is a common standard for clinical data analysis.
4. Reusability: Data should be reusable. We believe a software container is a viable method for addressing this principle, as it will quickly provide the means to explore shared data for secondary analyses.

To develop the software container, we use several open-source applications. First, to pre-package an operating system for use on any machine, we used two open source software virtualization solutions: VirtualBox[16] and Vagrant[17]. VirtualBox is a software virtualization environment that is designed to manage guest operating systems running within a primary host operating system. It's one of many technologies designed to perform this task, with other notable examples being Microsoft's Hyper-V and Dell's VMWare. The primary benefit of software virtualization is the ability to quickly and easily replicate the operating conditions of software without needing to replicate their expensive hardware environment. This allowed us to create a virtual computer, containing data and automated analysis scripts in a single container that can be run through another computer, regardless of the operating system (e.g. Microsoft Windows, Apple macOS, etc.). VirtualBox is the most widely used open source virtualization software and is used in health informatics for security and performance testing, but is being used more and more for the packaging of data and analysis pipelines for reuse [18, 19]. Vagrant is a virtualization management software designed to simplify the organization and description of virtual machine environments. Vagrant facilitates storing a robust description of the entire software environment needed to perform a given task. This software makes it easier for investigators to open the virtual machine and visualize results of their analysis. This software allowed us to encapsulate the dataset and the analytical software needed to perform analysis.

In these environments, the dataset is stored in a comma separated values (.csv) file, allowing easy access by analytical software. This standard file format is also readable into any spreadsheet program and requires minimal electronic storage space. This was desirable to limit the processing and memory overhead required by the virtual machine, allowing for more processing power to be devoted to the analysis engine.

Statistical analysis scripts were written in the R environment [20]. This is an open-source software commonly used for highlevel statistical analysis. Common analyses used by clinical investigators were re-created in this programming environment and packaged along with R version 3.3.2 and the clinical data inside of the virtual machine.

In the case of data sharing, the data and analysis scripts are stored in a folder along with a virtual machine description. When the machine is initiated using Vagrant, the dataset and

analysis scripts are loaded into the guest environment and the virtual machine is ready to perform the analysis and display results. A diagram of this structure is illustrated in **Figure 1**.

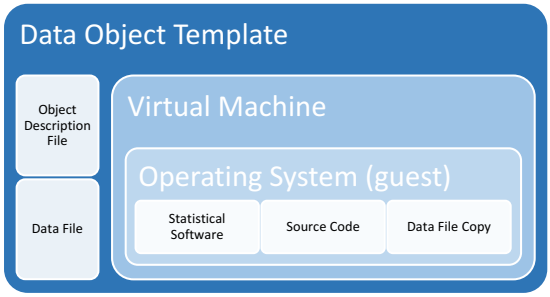

Fig. 1. Diagram of the data object template. Included with the data is the Object Description File, containing the configuration information needed to replicate the analysis environment, including the statistical software (R) and analysis source code.

The steps necessary to open the virtual machine and perform analysis are summarized as follows:

1. Ensure that Vagrant and VirtualBox are downloaded and installed on the local machine.
2. Download the Data Container and unzip into a directory (e.g. Computer desktop).
3. Double click on the startup file in the directory corresponding to your operating system (Microsoft Windows or Apple macOS).

The virtual machine initializes its startup sequence before loading the statistical programming environment and executing the packaged analysis script. At this point a user familiar with the R analysis software can explore the automatically loaded dataset and perform analysis.

This process was tested using the University of Louisville Pneumonia Study lactic acid dataset. The lactate study featured 3658 patients enrolled from June 1, 2014 to May 31, 2016. Lactate levels were associated with higher in-hospital mortality, with patients having ≥4 mmol/L of lactic acid having a nearly a 3-fold increase in odds of dying during hospitalization. The dataset is available for download[21]. The analysis scripts performed for this study included:

1. aggregate descriptive analyses (frequency with percent for categorical variables and mean with standard deviation for continuous variables),
2. bivariable comparisons of patient characteristics between those with and without lactate levels of ≥4 mmol/L using student's t-tests for continuous variables and Chi-squared tests for categorical variables,
3. univariable logistic regression for calculation of unadjusted odds ratios, and
4. multivariable logistic regression for calculation of adjusted odds ratios comparing the adjusted odds of in-hospital mortality for those with and without lactate levels of ≥4 mmol/L.

The variable names and descriptions are shown in **Table 1**. We record the time needed to display analysis results for this dataset on two different host platforms: Microsoft Windows and Apple macOS.

Table 1. Variable names and descriptions.

Variable	Description	Coding
age	Categorized Age in Years	
sex	Sex	1=Male 0=Female
nursinghome	Nursing Home Resident	1=Yes 0=No
neoplastic	History of Neoplastic Disease (past year)	1=Yes 0=No
liver	History of Liver Disease	1=Yes 0=No
chf	History of Congestive Heart Failure	1=Yes 0=No
cvd	History of Cerebrovascular Disease	1=Yes 0=No
renal	History of Renal Disease	1=Yes 0=No
mental	Altered Mental Status on admission	1=Yes 0=No
hr125	Heart Rate > 125 beats/minute	1=Yes 0=No
rr30	Respiratory Rate >30 breaths/minute	1=Yes 0=No
sbp90	Systolic Blood Pressure < 90 mmHg	1=Yes 0=No
templow	Temperature <35 degrees C	1=Yes 0=No
temphigh	Temperature >= 40 degrees C	1=Yes 0=No
ph735	Arterial pH <7.35	1=Yes 0=No
bun30	Blood Urea Nitrogen >=30 mg/dl	1=Yes 0=No
na130	Sodium <130 mmol/L	1=Yes 0=No
gluc250	Glucose >=250 mg/dl	1=Yes 0=No
hematocrit30	Hematocrit < 30%	1=Yes 0=No
pao260	Partial pressure of arterial O2 <60 mmHg	1=Yes 0=No
peffusion	Pleural effusion present	1=Yes 0=No
copd	History of COPD	1=Yes 0=No
diabetes	History of Diabetes	1=Yes 0=No
icudirect	Patient admitted directly to ICU	1=Yes 0=No
imv	Invasive mechanical ventilation on day 0	1=Yes 0=No
vaso	Vasopressors taken on day 0	1=Yes 0=No
psi4or5	Pneumonia Severity Risk Class IV or V	1=Yes 0=No
curb4or5	CURB-65 score 4 or 5	1=Yes 0=No
lactate	Lactate level	0= <2 mmol/L 1= 2-4 mmol/L 2= ≥4 mmol/L
ihm	In-hospital mortality	1=Yes 0=No
clinical_failure	Clinical Failure within two weeks of admission	1=Yes 0=No
los	Length of Stay (in days)	
los_yn	Patient discharged within 2 weeks	1=Yes 0=No
tcs	Time to clinical stability (in days)	
tcs_yn	Patient clinically stable within 1 week	1=Yes 0=No

Results

Host machine specifications and display times are shown in **Table 2**. The first startup time includes the time needed to download the initial virtual machine operating system, which will vary depending on many factors such as connection speed and network congestion. If the user shuts down the virtual machine after interacting with data, subsequent changes to the system will be much faster as shown in the subsequent startup time column. The large time difference in the two compared operating systems is due to the solid-state storage technology used in all new Apple computers, and not available in the Windows Server used in this study. A Windows system with solid state technology would have comparable startup times to the Apple system.

Table 2. Startup times for the software container on macOS and Windows platforms

Machine	First Startup Time	Subsequent Startup Times
Apple macOS MacBook Pro: 2Ghz i5 dual core 16GB Memory	5m:36s	0m:31s
Microsoft Windows Server 2012 Dell PowerEdge 610: 2.5Ghz i5 quad core 96GB Memory	7m:02s	2m:56s

The process of the virtual machine after downloading and installing is as follows, assuming the free Vagrant and VirtualBox software have also already been installed. First, the system will download a free Linux environment called Ubuntu [22]. After this has been downloaded, the virtual machine boots and starts downloading the current R software needed to perform analysis. Because R includes many different libraries needed to perform various analyses, this typically requires 2-3 minutes. Once the installation and configuration of R is complete, the user will be in the R command line environment and the system will have executed the output of the packaged study analysis. The results of the analysis is shown in **Figures 2 and 3**.

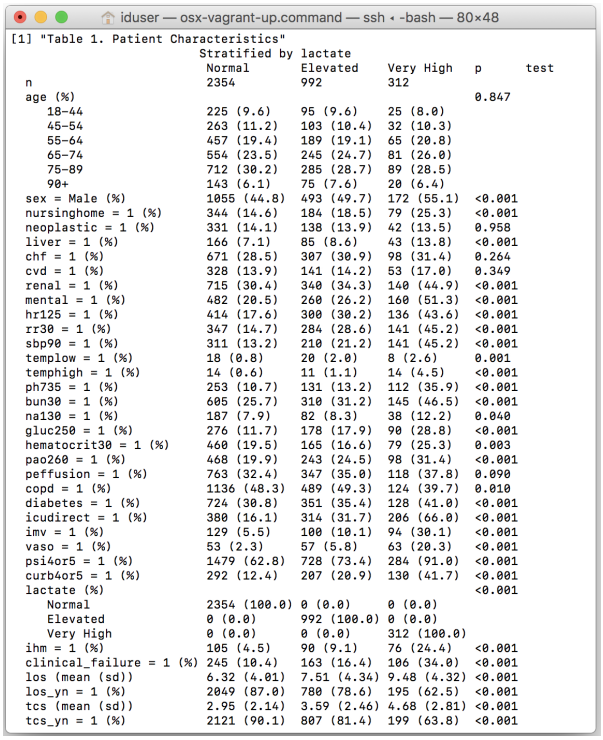

Fig. 2. Screenshot of the generated patient characteristics table for the University of Louisville Pneumonia Study Lactic Acid dataset.

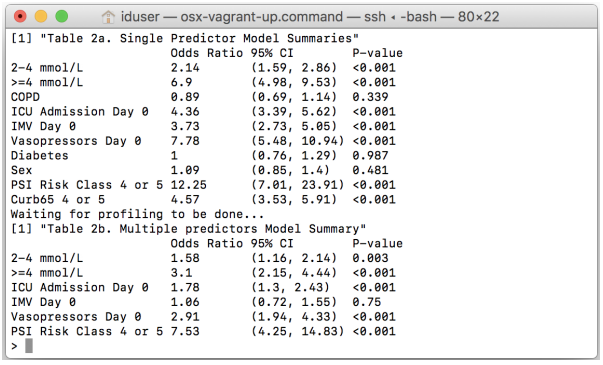

Fig. 3. A screenshot of the univariable and multivariable logistic regression output of the virtual machine.

Discussion

To our knowledge, this study was the first of its kind to create a pre-packaged software container for data sharing and automated statistical analysis of clinical research data. The open-source software used makes the container free and readily usable by all individuals with a computer and internet access. The container opens and installs rapidly, and provides automated output for results.

We believe that including the statistical software environment used to produce the results for a study dataset is an important contribution to data sharing and data authorship. We have developed a template for this type of data sharing for which the setup time needed to see and interact with results is negligible. Providing the details of an analysis exactly as they were performed is valuable to original study investigators and those wanting to perform secondary analyses.

The nature of data sharing is constantly changing and the most effective requirements are still an item of debate [23-27]. It is generally agreed that data sharing plans are beneficial to all research stakeholders, but the most cost-effective way to achieve data sharing is still unclear. The argument is often made that the only way to overcome the cost obstacles of data sharing requirements is to take advantage of a highly-centralized system with robust and standardized requirements for data and metadata. Systems like these are emerging and include: Yale Open Data Access (YODA) [28] and the Supporting Open Access for Researchers (SOAR) initiative [29] , but it is not clear how these data repositories will work together without an industry backed standard.

Another major concern for data sharing is fairness regarding differences in research infrastructure [30]. Countries and organizations with well-established research infrastructure are better equipped to discover knowledge from shared data sets. They will usually have strong analysis pipelines and trained biostatisticians and epidemiologists available to perform secondary analysis on collected and curated data. This may lead to the marginalization of smaller research groups who play an important role in collecting and providing data to the research community.

Further issues with data sharing include secondary investigators using shared data and publishing their results without acknowledgment of the initial research team. This issue often results in hesitation to share data. A more recent data sharing strategy suggests that authorship could be associated with a published dataset [31]. This allows the investigators and team responsible for collecting and curating a set of clinical data to publish it online in a public data repository. The data authors can then be referenced in publications by the original investigators themselves or by collaborators and secondary investigators. This allows original investigators to get the credit they deserve for studies that can be difficult to plan, set up, and manage. Many collaborative organizations are forming to try to mitigate the problem involving credit for secondary data use. The Community Acquired Pneumonia Organization [32] was established to facilitate advances in pneumonia research through collaboration and data sharing. Other groups include the Infectious Diseases Data Observatory [33], the Worldwide Anti-malarial Resistance Network [34], the National Surgical

Adjuvant Breast and Bowel Project [35] and many others. The benefits of such organizations are substantial and include development of better research questions and clear mission goals for produced research. One drawback is that while data will be consistent within such groups, a common data standard is needed to support true multidisciplinary collaboration.

There are several limitations to this study. First, The process we describe shifts some technical burden from a secondary investigator to the original investigators. There are many options available for packaging data objects and investigators will need to decide the most efficient means of data stewardship. Ultimately, we believe data stewardship and data authorship efforts will become formalized in an endorsed standard, making the creation process more streamlined and easy. Until that time, investigators should endeavor to follow FAIR principles to the best of their ability and make the data they share as accessible as possible. Second, the setup process will be specific to the type of operating system a secondary investigator is using. An effective container will support the three major operating systems, Windows, macOS, and Linux, but this greatly increases the work investment for investigators. Because of the similarities between macOS and Linux, supporting Windows and macOS is generally sufficient as they comprise 94.05 percent of the operating system market share in 2017[36]. Thirdly, it is always possible that secondary users will be able to misinterpret share data or the results of analysis. We have tried to mitigate this as much as possible by providing comments in the analysis software code and in the output of results.

Conclusion

We have described a data container capable of effectively sharing data along with the software code used to arrive at publishable results. In the future graphical plots should be added to data objects as they are an important part of understanding the results of research. We intend to develop software containers that quickly display graphical representations from within a data object. Possible means include packaging an interactive web environment with the data object or using the windowing interface of the host machine to display plots from the guest machine. Although the primary goal of this project was to outline how data can be shared and pre-packaged in an automated analysis environment, we believe this can also add to the transparency and reproducibility of clinical research findings through creation of software containers for results published in peer-reviewed journals or on clinicaltrials.gov. This increased transparency and facilitation of data sharing can enhance high quality research and translate into better patient care.

Acknowledgements: The authors wish to thank the University of Louisville Pneumonia Study Investigators and Collaborators for providing the data used to create the template for this project.

References

1. National Institutes of Health. NIH data sharing policy and implementation guidance. 2003. Available from: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
2. Guttmacher AE, Nabel EG, Collins FS. Why data-sharing policies matter. National Acad Sciences; 2009.
3. Hanson B, Sugden A, Alberts B. Making data maximally available. Science. 2011 Feb;331(6018):649–649.
4. Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE et al. Troubleshooting public data archiving: suggestions to increase participation. PLoS Biol. 2014 Jan;12(1):e1001779.
5. Lo B, DeMets DL. Incentives for clinical trialists to share data. N Engl J Med. 2016 Sep;375(12):1112–5.
6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3:160018.
7. Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I. Research objects: Towards exchange and reuse of digital knowledge 2010. https://doi.org/10.1038/npre.2010.4626.1.
8. De Roure D, Goble C, Stevens R. The design and realisation of the Virtual Research Environment for social sharing of workflows. Future Gener Comput Syst. 2009;25(5):561–7.
9. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E et al. Data sharing by scientists: practices and perceptions. PLoS One. 2011;6(6):e21101.
10. Krueger CW. Software reuse [CSUR]. ACM Comput Surv. 1992;24(2):131–83.
11. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5(3):299–314.
12. Pneumonia Research — UofL Pneumonia Study Group. 2017; Available from: http://louisville.edu/pneumonia-study
13. Zenodo - Research. Shared. 2017; Available from: https://zenodo.org/
14. CERN | Accelerating science. 2017; Available from: https://home.cern/.
15. DataCite. 2017; Locate, identify, and cite research data with the leading global provider of DOIs for research data. [cited 2017 September]. Available from: https://www.datacite.org/
16. Oracle VM VirtualBox. 2017 [cited 2017 September]; Available from: https://www.virtualbox.org/
17. Vagrant by HashiCorp. 2017 [cited 2017; Available from: https://www.vagrantup.com/index.html
18. Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M et al. Best practices in data analysis and sharing in neuroimaging using MRI. Nat Neurosci. 2017 Feb;20(3):299–303.
19. Dinov ID, Torri F, Macciardi F, Petrosyan P, Liu Z, Zamanyan A et al. Applications of the pipeline environment for visual informatics and genomics computations. BMC Bioinformatics. 2011 Jul;12(1):304–304.
20. The R Project for Statistical Computing. 2014 [cited 2014 April 11]; Available from: http://www.r-project.org/
21. Mattingly, W.A. and C. Sinclair, ul-research-support/CAPO-Lactic-Acid-2017: CAPO- Lactic-Acid-2017-Data-Object. 2017.
22. Ubuntu. 2017; Available from: https://www.ubuntu.com/
23. Haileamlak, M., et al., Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal Editors.

24. Longo DL, Drazen JM. Data sharing. Mass Medical Soc.; 2016.
25. Devereaux PJ, Guyatt G, Gerstein H, Connolly S, Yusuf S; International Consortium of Investigators for Fairness in Trial Data Sharing. Toward fairness in data sharing. N Engl J Med. 2016 Aug;375(5):405–7.
26. Rockhold F, Nisen P, Freeman A. Data sharing at a crossroads. N Engl J Med. 2016 Sep;375(12):1115–7.
27. Rosenbaum L. Bridging the Data-Sharing Divide - Seeing the Devil in the Details, Not the Other Camp. N Engl J Med. 2017 Jun;376(23):2201–3.
28. The YODA Project. 2017; Available from: http://yoda.yale.edu/
29. Navar AM, Pencina MJ, Rymer JA, Louzao DM, Peterson ED. Use of open access platforms for clinical trial data. JAMA. 2016 Mar;315(12):1283–4.
30. Merson L, Gaye O, Guerin PJ. Avoiding data dumpsters—toward equitable and useful data sharing. N Engl J Med. 2016 Jun;374(25):2414–5.
31. Bierer BE, Crosas M, Pierce HH. Data Authorship as an Incentive to Data Sharing. The New England journal of medicine. 2017 Apr 27;376(17):1684-7.
32. Community-Acquired Pneumonia Organization. 2017; Available from: http://caposite.com/
33. Infectious Diseases Data Observatory | Home. 2017; Available from: https://www.iddo.org/
34. Worldwide Antimalarial Resistance Network | Home. 2017; Available from: http://www.wwarn.org/
35. National Surgical Adjuvant Breast and Bowel Project (NSABP). 2017; Available from: http://www.nsabp.pitt.edu/
36. Operating system market share. 2017; Available from: https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0