Appendix 1. Glossary of Advanced Data Methods in Biomedical Research

Adaptive (Bayesian) Study Design:

A study that is designed to be adapted, usually at some interim point in the study, based upon a prospective adaptation plan utilizing advanced statistical techniques, usually Bayesian. This allows a study to examine previously unforeseen important variables midway through the study's lifetime, compromising the traditional hypothesis (p-value) based statistics. Instead relatively reliable statistical inference is used to create a practical approach with active selection of samples. This design is often useful for small sample sizes with a large number of variables.

Principal Component Analysis (PCA):

PCA uses an orthogonal transformation to convert correlated variables into a set of linearly uncorrelated variables. The new variables, called principal components, are ordered according to their variance. This allows for dimension reduction through omission of principal components that have low variance. A recent application of PCA in obstetrics was the identification of the alterations in the NS1 as a key to understanding how Zika Virus became such a threat to humans in Asia and the Americas.

Bootstrapping:

A type of statistical test that requires sampling, with replacement, on a sample dataset in order to determine aspects of the larger population (from which the sample dataset was sampled). Performed a large number of times, this can, in select cases, project additional statistical information on the original dataset without having to collect additional samples.

Cross-validation:

Cross-validation is a method for improving model prediction performance by partitioning a set of data into a training set and a validation set important for reducing Type III Error (correctly rejecting the null hypothesis for the wrong reason). This technique is used in oncology in optimal biomarker selection.

Data Cleansing:

Data cleansing is a process by which data are scrubbed to improve the validity of analytical work performed on a particular data set. This may include the removal of duplicate files, filtering for incomplete or inaccurate data, and the detection and deletion of corrupt files.

Macedonia CR and Johnson CT. Advanced research and data methods in women's health: big data analytics, adaptive studies, and the road ahead. Obstet Gynecol 2017; 129.

The authors provided this information as a supplement to their article.

Data Mining:

See “Predictive Analytics”

Data Visualization:

A branch of computer science that involves the presentation of data in some form of a visual format rather than purely digital or symbolic form. It may take the form of a graph, chart, or artistic rendering in addition to numbers or symbols. The main purpose of data visualization is to make complicated data more accessible to the receiver of the information.

Deep Learning

Deep Learning (Deep Neural Network analytics) is a form of Machine Learning utilizing neural networks but goes beyond traditional neural network techniques. Large set of biological data by their very nature contain structure. Deep learning systems are ones capable of both learning the structure inherent in the data and then applying the learned structures to further learn from the data at a deeper level.

Dimensionality Reduction: Linear and Nonlinear Dimensionality Reduction

Dimensionality Reduction is a statistical technique that reduces the number of random variables under consideration. It is useful in situations where there is a data set with a large number of dimensions, but the principal variables are contained within a few dimensions. This suggests that some dimensions are insignificant and can be omitted. This has been used in epidemiology, genomics, electrophysiology, and neuroscience. The most widely used algorithms are variations of Principal Component Analysis or PCA.

Non-Linear Dimensionality Reduction, as the name would imply, attempts to reduce the dimensions of a dataset without assuming that variables scale linearly. Data may lie within a non-linear “manifold”, and this manifold may be approximated through a number of manifold learning algorithms. The two most widely used algorithms are Laplacian eigenmaps and Self-organizing maps. These techniques are widely used in such applications as facial recognition and computer aided mammography.

Macedonia CR and Johnson CT. Advanced research and data methods in women’s health: big data analytics, adaptive studies, and the road ahead. Obstet Gynecol 2017; 129.

The authors provided this information as a supplement to their article.

Eigenvalues and Eigenvectors:

Eigenvalues: a special set of scalars associated with certain linear systems of equations.[i] The decomposition of a square matrix $\mathbf{A}$ into eigenvalues and **eigenvectors** (the vectors associated with eigenvalues) is known in this work as eigen decomposition. Eigenvalues are represented by λ and eigenvectors by x. $\mathbf{A}x = \lambda x$ with $x \neq 0$.

By way of analogy, one may think of a dataset as being like a very large herd of sheep invaded by a wolf. You may not be able to see the wolf if you are far away but you can surmise its location anyway because the sheep are all facing or running away from the location of the wolf. So too can too can the strongest influencing variable in a large data set drive the other variable in such a way that the mass of variables point back to (vector toward) this variable, perhaps not individually, but in the summation of all of their changes or movements. In mathematical terms, this is an eigenvalue or eigenvector.

Family based study design (triad association studies, subset of a case control study)

Studies of the influence of genes by studying families. The most popular of these studies to date have been twin studies but family based studies may involve triads (two affected family members clustered with one who is not) or other forms of linkage analysis.

Imputation-based association analysis

Used in genome wide association studies, this computational technique finds correlations between highly dense reference samples and far less dense experimental samples. This technique leverages mathematical principles found in information theory and cryptography.

Genome Wide Association Study (GWAS)

A study involving the sequenced genomes of many individuals and mathematical tools to look for variations in genetic markers with respect to correlations in phenotypic variability, traits, or disease. GWAS studies looks at the sequenced genome as the state variable, and physical traits as the output variable. They then sample a large number of people to determine if correlations exist between unique aspects of a person's genome and the physical traits they display.

Logistic Regression or Multivariate Logistic Regression:

Is a statistical model used to determine the probability of an outcome, given a set of independent and corresponding dependent variables. The probability of the outcome is the dependent variable and the influencers of that outcome are the independent variables.

Macedonia CR and Johnson CT. Advanced research and data methods in women's health: big data analytics, adaptive studies, and the road ahead. Obstet Gynecol 2017; 129.

The authors provided this information as a supplement to their article.

Knowledge Discovery in Databases (KDD)

See "Predictive Analytics"

Machine Learning (Supervised and Unsupervised):

Is a discipline of computer science concentrated on using artificial intelligence to analyze information, find patterns or features, and make novel predictions without requiring explicit programing to do so. Prior training datasets may be used to guide the program to look for certain patterns of interest (supervised), or analysis may be blind to any human input, extracting only the patterns statistically determined to be the most significant (unsupervised).

Manifold Learning

Manifold Learning (non-linear dimensionality reduction) pursues the goal of embedding data that originally lie in high dimensional space in lower dimensional space, while preserving characteristic properties. The most widely used algorithm for manifold learning is kernel PCA.

Markov Chain Monte Carlo (MCMC) and Hidden Markov Modeling (HMM)

A Markov Chain is a collection of random variables, output from certain state spaces, whose future values are only dependent on current state and not on prior values. Monte Carlo simulators are a class of algorithms that generate random numbers for use in problems requiring a probabilistic interpretation, relying on a very large number of simulations to determine the optimum result. This is a highly computationally intensive technique originally developed for nuclear weapons research, now used in epidemiology, cancer, and genomics research. Markov Chain Monte Carlo is a Monte Carlo approach for sampling from a probability distribution by constructing a suitable Markov chain that has the desired distribution as its stationary distribution, and then simulating the Markov chain. Hidden Markov Model is a Markov model whose states are hidden and one can observe only outputs which are functions of the state. Hidden Markov modeling looks at these outputs from an underlying Markov chain (ie memoryless and output from distinct state spaces), and attempts to determine the state from which each variable was derived through analysis of the output variables.

Metadata:

Literally means "data about data." Metadata is a class of data that describes collections of other data either individually or in groups, and can include intricacies such as the time of collection, operator

of collection, collection protocol, and manufacturer of collection instruments. One area where this is readily used in women's health research is in imaging research. Radiological Information Systems such as ultrasound reporting systems store image files appended with additional data to make the images searchable and relatable to other studies. This data concerning the content of other data is called metadata.

Neural Networks (or Artificial Neural Networks)

Neural Networks are networks designed with an architecture mimicking, in some form, the networks of human neurons. These networks not only have the interconnectedness of neuronal networks but also have the ability to learn. Because of these features these networks are tolerant to processing information that is incomplete they can compute approximations with varying degrees of accuracy.

p-Hacking

Also known as "data dredging" is the abuse of statistical tools to sift through data to find correlations meeting a statistical significance threshold ($p < 0.05$) thus finding a "meaningful" relationship when none actually exists. There is always a probability that a statistically significant relationship can be found by random chance between any two dynamic random variables, but that does not mean a true correlation exists. Proper application of data mining tools involves the application of proper controls to reduce bias, such as sampling replicates and cross-validation.

PICO/PECO Process (Problem/Population, Intervention/Exposure, Control, Outcome)

A rubric for evidence based study design useful in building statistically valid experimental, observational or cohort studies with an emphasis on minimizing investigator bias.

Predictive Analytics:

A field of computer science originating in the 1960's, also known as "knowledge discovery in data" (KDD) or "data mining". This is the field of computer science directed toward the discovery of patterns within systems data sets to derive an understanding of the behavior of the studied system and in some cases make predictions about the future. Data collection and storage has increased rapidly in recent years, resulting in more data than can be analyzed by the collecting persons. Predictive analytic algorithms attempts to sift through these under-explored datasets and extract meaningful correlations or patterns.

Support Vector Machine (SVM)

Macedonia CR and Johnson CT. Advanced research and data methods in women's health: big data analytics, adaptive studies, and the road ahead. Obstet Gynecol 2017; 129.

The authors provided this information as a supplement to their article.

Widely used in biology, this is a type of supervised machine learning is used in pattern recognition where there is an assumption of separability. This method finds within large data separating "hyperplanes" and Is used in a variety of biomedical applications including protein identification and diabetes prediction. It should be noted that SVM classification systems ae supervised, while PCA are unsupervised.

[i] Marcus M., Minc H., *Introduction to Linear Algebra*, New York: Dover, p 145, 1988.

Macedonia CR and Johnson CT. Advanced research and data methods in women's health: big data analytics, adaptive studies, and the road ahead. Obstet Gynecol 2017; 129.

The authors provided this information as a supplement to their article.

