公告【彈性密集課程旁聽登記】 115.10.29

授課教師：石瑜 客座特聘講座

Yu Shyr

Chair, Department of Biostatistics
Harold L. Moses Chair in Cancer Research
Director, Vanderbilt Center for Quantitative Sciences (CQS)
Director, Vanderbilt Technologies for Advanced Genomics Analysis and Research Design (VANGARD)
Professor of Biostatistics, Biomedical Informatics, and Health Policy

課程名稱：「**Clinical Research in the Data Science Era: Design, Bias, Statistical Methods, and Machine Learning**」

（資料科學時代的臨床研究：設計、偏差、統計方法與機器學習）（課程大綱參見附件）

開課系所：統計與資料科學學系

上課教室：醫學院第四講堂

上課時間：

1. 115.12.19(六) 2:30 PM – 5:30 PM (Lab: 4:30 PM – 5:30 PM)
2. 115.12.20(日) 9:00 AM – 12:00 PM、1:00 PM – 4:00 PM (Lab: 2:30 PM – 4:00 PM)
3. 115.12.21(一) 6:00 PM – 9:00 PM (Lab: 8:00 PM – 9:00 PM)
4. 115.12.22(二) 6:00 PM – 9:00 PM (Lab: 8:00 PM – 9:00 PM)

即日起接受旁聽登記，登記方式：

1. 開放 google 表單供登記旁聽：https://forms.gle/aY5cJSsWrE4PEBTq5
2. 限 50 人，登記額滿截止。

＊不提供線上旁聽。

連絡人：

吳季靜助教/統計與資料科學學系暨數據科學研究所

電話：06-2757575-53633

Email：wucc@ncku.edu.tw

Syllabus

Clinical Research in the Data Science Era: Design, Bias, Statistical Methods, and Machine Learning

資料科學時代的臨床研究：設計、偏差、統計方法與機器學習

National Cheng Kung University
December 19-22, 2026
Syllabus

Instructors Yu Shyr, Ph.D.
Chairman, Department of Biostatistics
Director, Vanderbilt Center for Quantitative Sciences
Harold L. Moses Chair in Cancer Research
Professor of Biostatistics, Biomedical Informatics, and Health Policy
Vanderbilt University Medical Center
E-mail: yu.shyr@vanderbilt.edu

Chih-Yuan Hsu, Ph.D.
Research Assistant Professor
Department of Biostatistics
Vanderbilt University Medical Center
E-mail: chih-yuan.hsu@vumc.org

This course introduces key concepts in applied artificial intelligence, machine learning, and modern biostatistics for biomedical and precision medicine research. Contemporary biomedical studies increasingly integrate molecular profiles, clinical characteristics, imaging, electronic health records, and other high-dimensional data to improve disease prevention, diagnosis, prognosis, and treatment selection. Technologies such as single-cell next-generation sequencing, spatial transcriptomics, circulating tumor DNA, microbiomics, radiomics, and other high-throughput platforms have greatly expanded the amount and complexity of biomedical data available for research.

The rapid growth of large-scale biomedical and real-world data creates substantial opportunities, but also important methodological challenges. More data and more sophisticated algorithms do not necessarily produce more reliable evidence. Poor study design, selection bias, confounding, measurement bias, missing data, inappropriate analytic methods, and overfitting can substantially compromise both traditional statistical analyses and machine-learning studies.

This course will therefore emphasize the integration of rigorous study design, bias assessment, modern statistical methodology, and machine learning. Particular attention will be given to common methodological problems encountered in published biomedical research and to practical approaches for improving reproducibility, validity, and clinical relevance. The course will also introduce emerging methods for prediction, causal inference, individualized treatment effects, and evaluation of machine-learning models.

I. Design, Bias, and Conduct of Modern Biomedical Research

This section will address fundamental principles for designing rigorous biomedical studies in the era of AI and large-scale data. Topics will include defining the primary scientific objective and endpoint, selecting appropriate study designs, determining sample size and statistical precision, and distinguishing prediction, association, and causal inference.

A major emphasis will be placed on **bias**. Students will learn to recognize selection bias, confounding, measurement bias, immortal-time bias, informative missingness, spectrum bias, data leakage, and other sources of bias that commonly arise in observational studies, real-world data analyses, and AI-based research. Strategies for preventing or reducing these biases through study design and statistical analysis will be discussed using examples from biomedical and clinical research.

Modern approaches for creating comparable treatment or exposure groups will also be introduced. These include propensity-score matching, optimal matching, full matching, inverse probability weighting, overlap weighting, coarsened exact matching, and other balancing approaches. Particular attention will be given to evaluating covariate balance rather than relying solely on traditional significance testing of baseline characteristics.

The course will also address missing data, internal and external validation, model transportability, reproducibility, and appropriate separation of training, validation, and test datasets in machine-learning studies.

II. Statistical Analysis, Machine Learning, and Causal Data Science

This section will cover statistical methods for common biomedical research designs, including randomized clinical trials, Bayesian clinical trials, cohort studies, case-control studies, cross-sectional studies, and longitudinal or repeated-measures studies. Methods for continuous, categorical, count, and time-to-event outcomes will be reviewed with an emphasis on selecting methods that appropriately address the scientific objective and study design.

Modern regression and machine-learning approaches will be introduced, including penalized regression, random forests, boosting, bagging, and stacking. The course will emphasize not only how these methods are implemented, but also how their performance should be evaluated and interpreted in biomedical applications.

Prediction-model evaluation will include discrimination, calibration, sensitivity, specificity, predictive values, ROC-based measures, and the F1 score, with discussion of when each metric is appropriate and when commonly used measures may be misleading.

The course will also introduce causal inference and individualized treatment-effect analysis. Topics will include propensity-score methods, weighting approaches, treatment-effect heterogeneity, and methods for evaluating whether predictive algorithms can identify patients who benefit differently from alternative treatments. Modern evaluation approaches such as uplift modeling and Qini statistics will be discussed in this context.

Throughout the course, published biomedical studies will be used to illustrate common statistical and methodological errors, including inappropriate subgroup analyses, failure to account for multiplicity, misuse of predictive terminology, inadequate validation, overfitting, inappropriate handling of missing data, and failure to distinguish association from prediction or causation.

The overall goal of the course is to help students develop the ability to critically evaluate modern biomedical research and to apply rigorous statistical and data-science methods in an era increasingly shaped by artificial intelligence, machine learning, real-world data, and precision medicine.

III Topics:

Date	Topic
Dec 19 (Saturday) 2:30 PM – 5:30 PM (Lab: 4:30 PM – 5:30 PM)	**Foundations of Modern Biomedical and Clinical Data Science** Overview of AI, machine learning, and modern biostatistics in biomedical research From scientific question to study design: association, prediction, and causal inference Major sources of bias in biomedical research • Selection bias • Confounding • Measurement and information bias • Immortal-time bias • Data leakage and other biases in AI and machine-learning studies Primary objectives, endpoints, and study design Sample size determination, power, and precision analysis Introduction to Bayesian methods and modern clinical trial designs **Common fatal design and statistical flaws in published biomedical research** **Lab 1:** Bayesian computation and Markov Chain Monte Carlo

Dec 20 (Sunday) 9:00 AM – 12:00 PM 1:00 PM – 4:00 PM (Lab: 2:30 PM – 4:00 PM)	**Statistical Modeling for Biomedical Research** Univariate data analysis: parametric and nonparametric methods Regression modeling for biomedical outcomes • Linear regression • Logistic regression • Poisson and negative binomial regression • Models for ordinal outcomes Model specification, variable selection, and interpretation Missing data mechanisms and principles of missing data analysis Multiple imputation and sensitivity analysis **Common statistical mistakes in biomedical data analysis** **Lab 2:** Generalized linear models, missing-data imputation, and variable selection
Dec 21 (Monday) 6:00 PM – 9:00 PM (Lab: 8:00 PM – 9:00 PM)	**Longitudinal, Survival, and Causal Inference Methods** Analysis of longitudinal and repeated-measures data • Mixed-effects models • Generalized estimating equations Time-to-event analysis and Cox proportional hazards models Causal inference using observational and real-world data Propensity-score methods and modern approaches to covariate balance • Propensity-score matching • Optimal and full matching • Coarsened exact matching • Inverse probability of treatment weighting • Overlap weighting Assessing covariate balance and diagnosing residual confounding **Bias, target populations, and interpretation of treatment effects** **Lab 3:** Mixed-effects models, GEE, and propensity-score methods
Dec 22 (Tuesday) 6:00 PM – 9:00 PM (Lab: 8:00 PM – 9:00 PM)	**Machine Learning and Modern Clinical Data Science** Machine learning for high-dimensional biomedical data Penalized regression • Ridge regression • LASSO • Elastic Net Ensemble learning • Bagging • Random forests

	• Boosting • Stacking Training, validation, testing, overfitting, and data leakage Evaluation of prediction and classification models • Sensitivity, specificity, and predictive values • ROC curves and AUC • Precision, recall, and **F1 score** • Calibration and clinical utility Treatment-effect heterogeneity and **individualized treatment effects** **Uplift modeling and Qini statistics** for evaluating treatment-selection algorithms Integration of statistical inference and machine learning in precision medicine **Lab 4:** Machine-learning applications using real biomedical data

IV. Software

1. PS Sample Size and Power Calculations (developed by William Dupont, free download).
2. R

V. Grade

Lab report (35%); Final (65%)