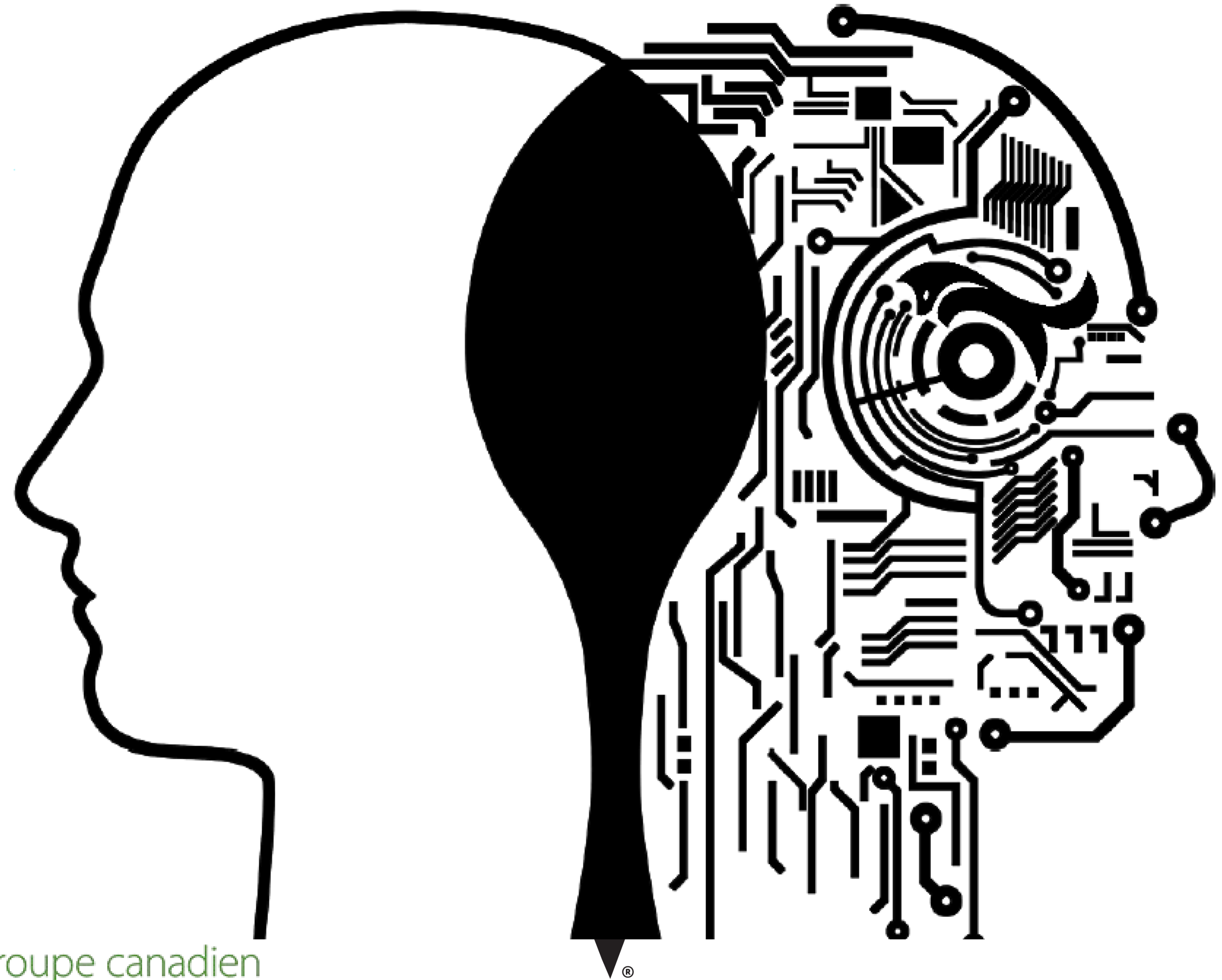
Addressing Bias in AI-Driven Medical Imaging: Pitfalls and Best Practices

Amber Simpson, PhD
Canada Research Chair in Biomedical Computing and Informatics
Associate Professor, DBMS/School of Computing
Director, Centre for Health Innovation
Senior Investigator, Canadian Cancer Trials Group

Affiliate Member, Vector Institute for AI

amber.simpson@queensu.ca
simpsonlab.org
@profsimsim



Vector Institute | Institut Vecteur

Centre for Health Innovation | Queen's University | Kingston Health Sciences Centre

Canadian Cancer Trials Group | Groupe canadien des essais sur le cancer

Nothing to disclose

My Perspective

Memorial Sloan Kettering Cancer Center

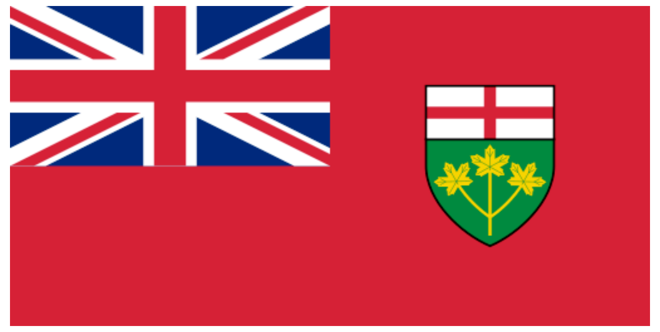
- High volume cancer centre
- Rich, deep data on cancer patients: genomics, imaging, etc.
- Single institution closed data; difficult to share

Canadian Cancer Trials Group / Groupe canadien des essais sur le cancer

- Coordinates and centralizes cancer trials across Canada
- Partners with EORTC and NCI
- Clinical trial specimen and data are centralized
- Data science platform for research

OHDP | ONTARIO HEALTH DATA PLATFORM COVID-19

- Population-level data in a single-payer health system
- Admin data (billing, lab, etc)
- 14 million Ontarians
- Data platform for research



NIH NATIONAL CANCER INSTITUTE Imaging Data Commons

- Cloud-based repository of publicly available cancer imaging data
- Analysis and exploration tools and resources
- Imaging and clinical data

Imaging AI Models

Images are data
Imaging AI models have the potential to transform care.

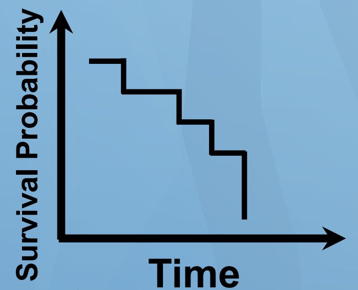
Promising results
Predictive and prognostic biomarkers
Quantitative and non-invasive
Inexpensive (standard of care imaging)

Bias and pitfalls
Very few imaging signatures have been clinically adopted.

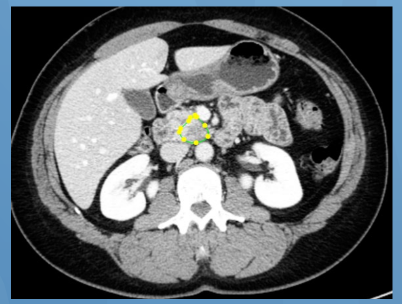
There are common sources of bias in imaging AI studies.

We suggest ways to address these.

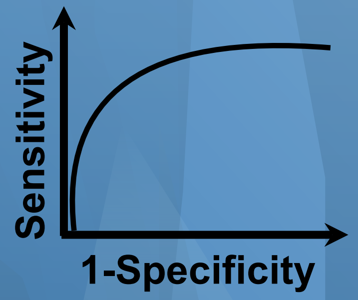
Study design
Incorporation bias
Verification bias
Spectrum bias



Prediction model building
Overfitting bias
Incorrect inference



Imaging data acquisition
Scanner & segmentation variability
Image analysis variability
Operator & software variability



Model evaluation
Optimistic performance bias

Adapted from: C. Moskowitz, M. Welch, B. Kurland, A. L. Simpson, Considerations in the design, conduct, and reporting of radiomic analyses, Radiology, 2022

Sources of Statistical Bias and Variability

Table 1: Frequent Sources of Variability and Bias in Radiomic Analyses

Type	Description
Study design	
Incorporation bias (23–25)	The outcome uses information from the images Example: Predicting the outcome from CT ima CT imaging
Verification bias (15,26)	Analysis only includes cases where the outcome population of interest Example: Only including patients with biopsies imaging
Spectrum bias (23)	Study data are not fully representative of the po Example: Model developed using only extren
Image acquisition and processing	
Scanner variability*	Scanner manufacturer, model, and/or calibratio Example: CT images obtained using different k reconstruction algorithms result in poor repr
Image analysis variability*	Variability arises when different filters, threshol Example: Texture features vary based on the dis of bins) (77)
Operator variability*	Manual or semiautomated segmentation affects Example: Inter- and intraoperator variability ex by the disease site (78) and existing clinical c
Software variability*	Feature measurement of the same region of inter Example: Hand-engineered features calculated o of the same software, can have different values
Statistical analysis	
Bias due to overfitting (65)	Model captures spurious associations in the trai replicated in similar data sets Example: A model captures random variation (but does not work well in independent valid
Optimistic performance bias (43,81)	Evaluating the algorithm on the same data that Example: A model is developed to optimize per assessed using both training and validation d
Bias from exclusion of indeterminate or missing feature data	Ignoring images with missing feature measurem the features and the algorithm's performance, (15,59) Example: Texture analysis requires a sufficient n multiple tumors, small tumors cannot be me

Table 2: Methods to Prevent Sources of Variability and Bias in Radiomic Analyses

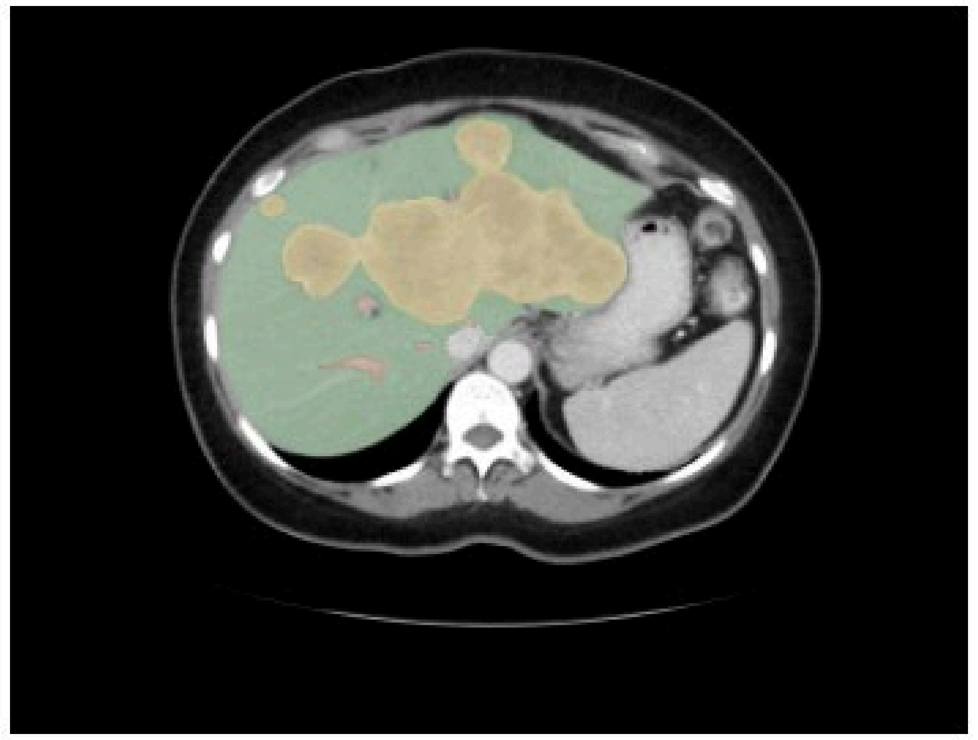
Type	Prevention
Study design	
Incorporation bias (23–25)	Exclude the index images and imaging modality from the definition of the outcome
Verification bias (15,26)	1. Ensure the outcome is evaluated for all patients, or 2. Ascertain the outcome on a random sample of patients, and/or 3. When analyzing data, use statistical methods developed for correction of verification bias (22,28–31)
Spectrum bias (23)	Ensure study data are generalizable to the population of interest; perform external validation on different data sets within the population of interest
Image acquisition and processing	
Scanner variability* Image analysis variability* Operator variability*	There are no prevention methods for these issues; these are open areas of research. We suggest the following: 1. Design controlled experiments to fully characterize the variability 2. Control for scanner effects when analyzing the data 3. Reduce and correct the variability to ensure results are generalizable 4. Validate models on another institution's data
Software variability*	1. Use consistent software pipelines 2. Use open-source software or release source code publicly 3. Adopt standardized feature sets (eg, Image Biomarker Standardization Initiative [52]) 4. Benchmark comparison, if not using the standard
Statistical analysis	
Bias due to overfitting (65)	1. Reduce the number of imaging features being studied 2. Ensure sample sizes are large enough to preclude spurious correlation, including in subgroups of interest 3. Use a resampling method such as cross-validation 4. Use a penalized regression method to build the algorithm 5. Evaluate the algorithm on an independent data set
Optimistic performance bias (43,81)	1. Use an entirely independent data set to evaluate the algorithm 2. In the absence of independent validation data, use cross-validation
Bias from exclusion of indeterminate or missing feature data	1. Disclose characteristics and amount of indeterminate and missing data 2. Evaluate associations among missingness and values of the outcome and other features 3. Perform sensitivity analyses treating missing features as positive and then as negative for binary features

C. Moskowitz, M. Welch, B. Kurland, A. L. Simpson, Considerations in the design, conduct, and reporting of radiomic analyses, Radiology, 2022

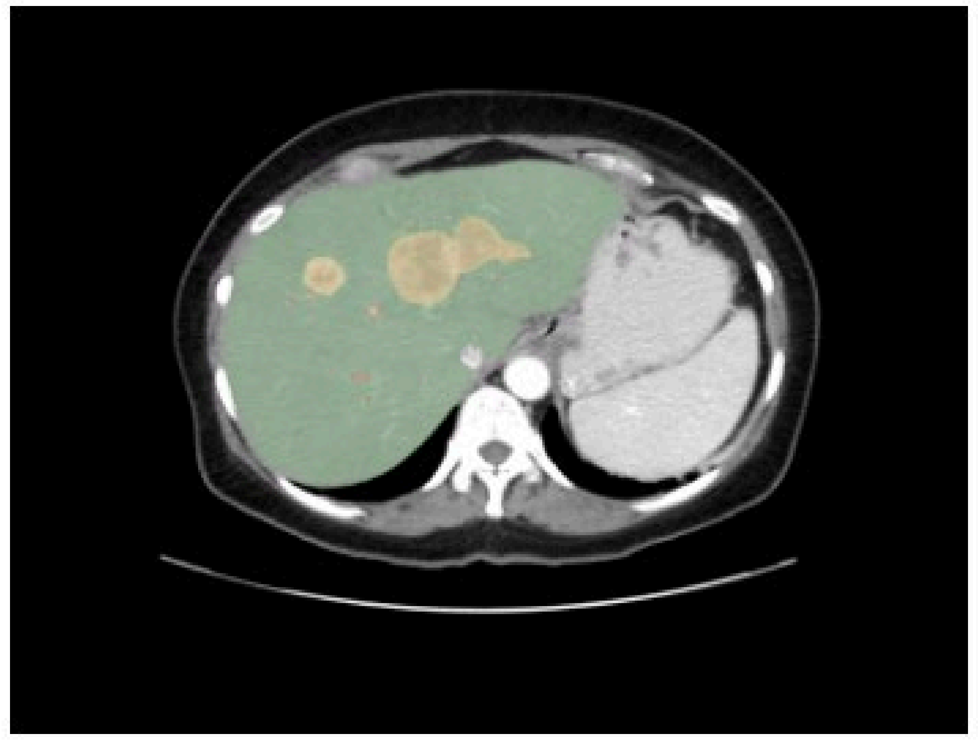
Sources of Statistical Bias & Variability - Study Design

Incorporation bias: The outcome uses information from the predictors

- Predicting response from CT images where response = diameter change



Pre treatment



Post treatment

Predicted = Diameter Change (RECIST)

- Gold standard response is pathology (requires biopsy)
- Bronze standard is radiology (best we can do despite weaknesses)

Example from my lab: M. Piliposyan et al., SPIE Medical Imaging 2025.

Sources of Statistical Bias & Variability - Study Design

Verification bias: Analysis only includes cases where the outcome is ascertained, which is a non-representative subset of the population of interest

- Example: Only including patients with biopsies where the decision to biopsy is determined based on imaging
- Risks underestimating the number of false negatives and thus may overestimate the sensitivity of a new test

EBM Learning

Verification bias FREE

Jack W O'Sullivan [1], Amitava Banerjee [2], Carl Heneghan [3], Annette Pluddemann [3]
Correspondence to Dr Jack W O'Sullivan, Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, OX1 2JD, UK; jack.osullivan@phc.ox.ac.uk

https://doi.org/10.1136/bmjebm-2018-110919

Sources of Statistical Bias & Variability - Study Design

Spectrum bias: Study data are not fully representative of the population of interest

- The performance of a diagnostic test may vary in different clinical settings because each setting has a different mix of patients
- Example: Model developed using only extreme cases (e.g. very sick and/or very healthy individuals)
- Occurs when assays are expensive
- ML papers that formulate harder problems as classification problems (survival, regression, etc)

Sources of Statistical Bias & Variability - Study Design

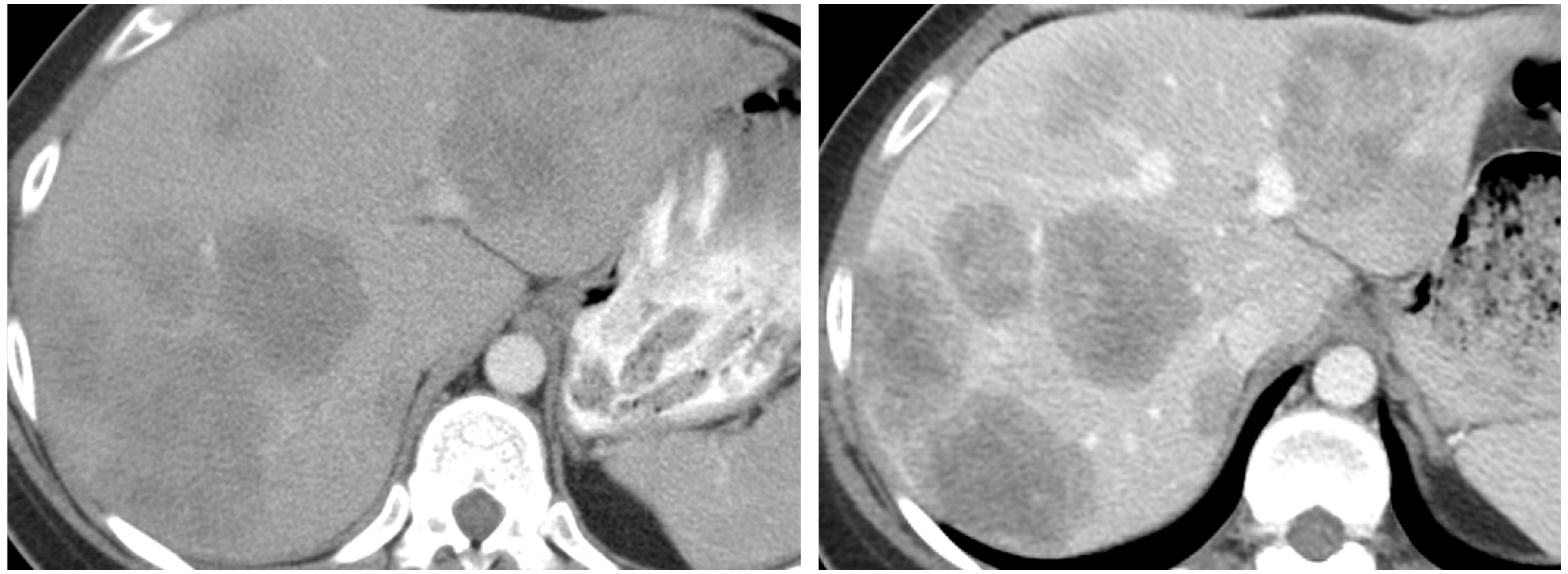
Samplin
member y
than oth

- I
  a
- I

Sources of Statistical Bias & Variability - Acquisition

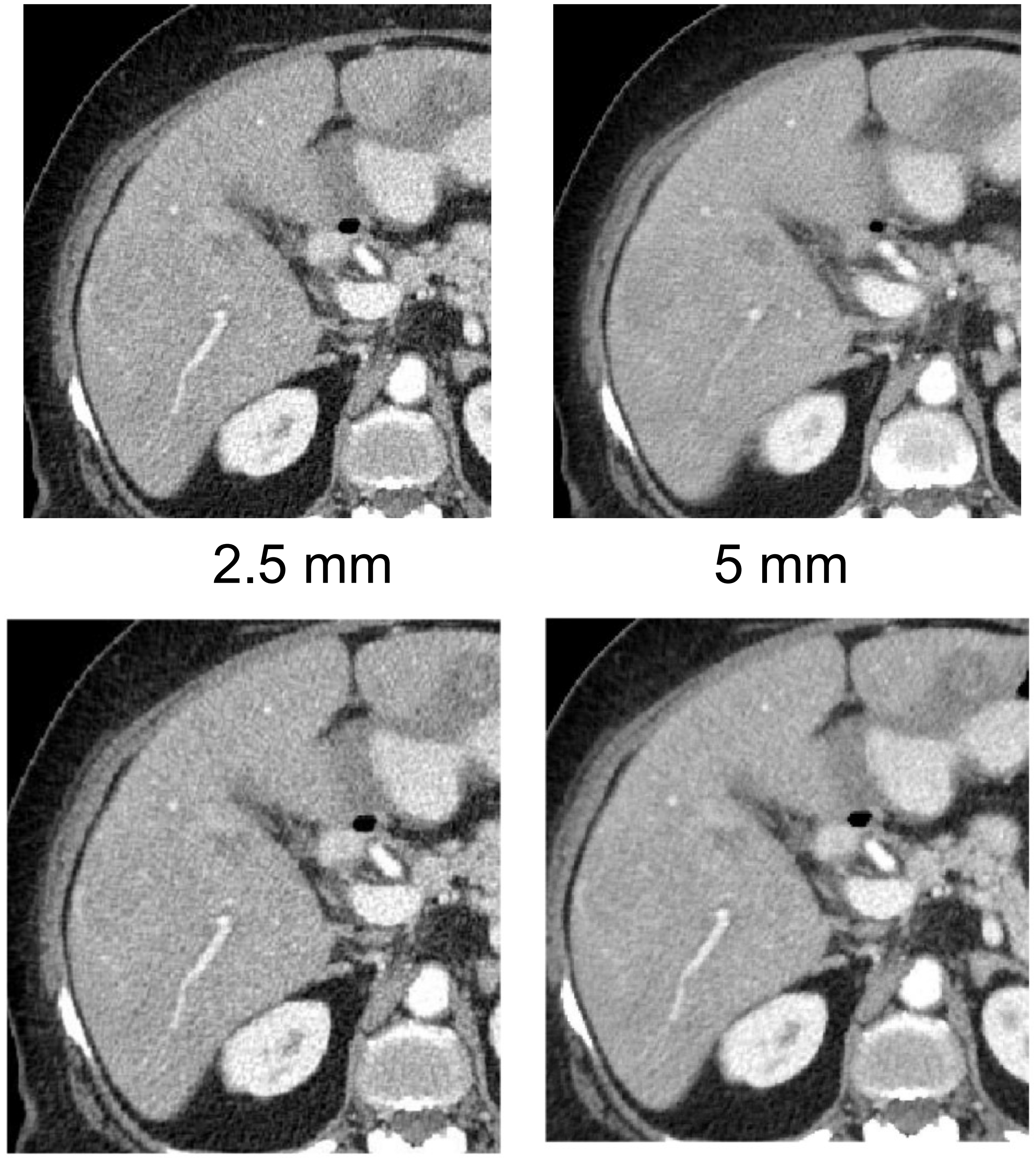
Device variability: Device manufacturer, model, and/or calibration differences influence appearance

- Example: CT images collected with different protocols and dose reduction strategies



Institution #1
June 19

Institution #2
June 29



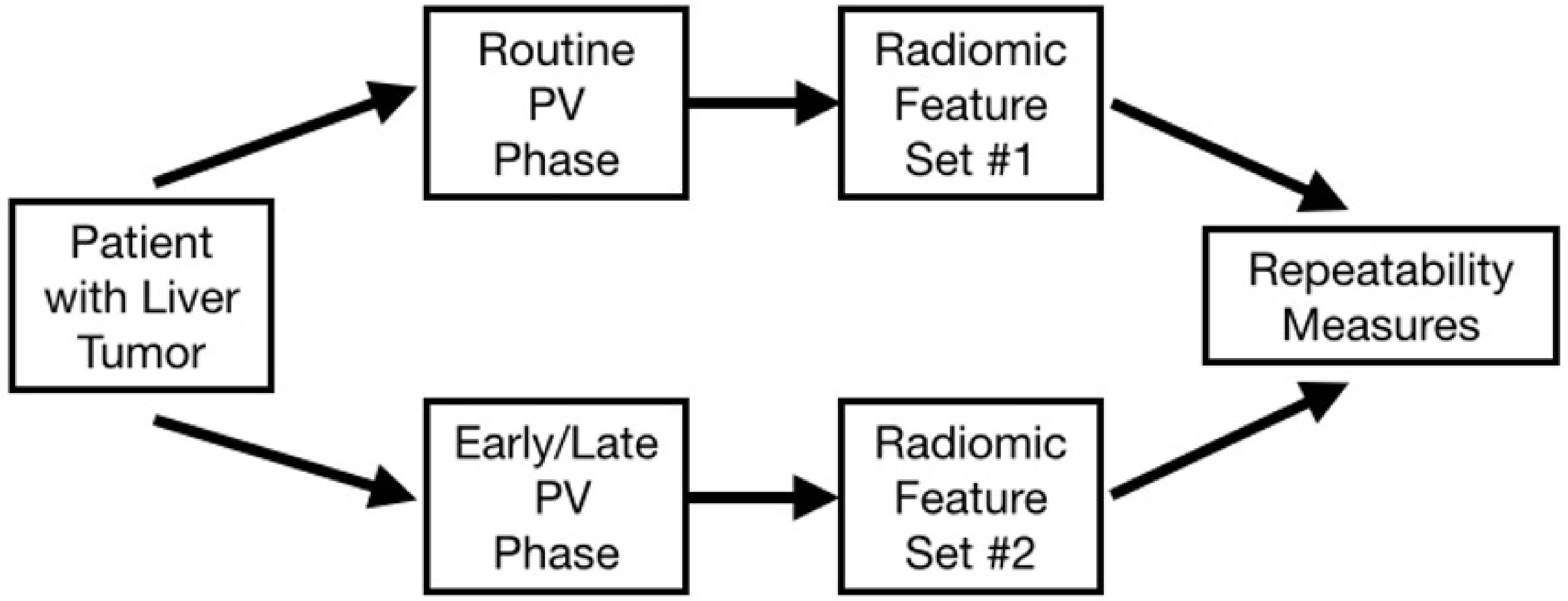
2.5 mm

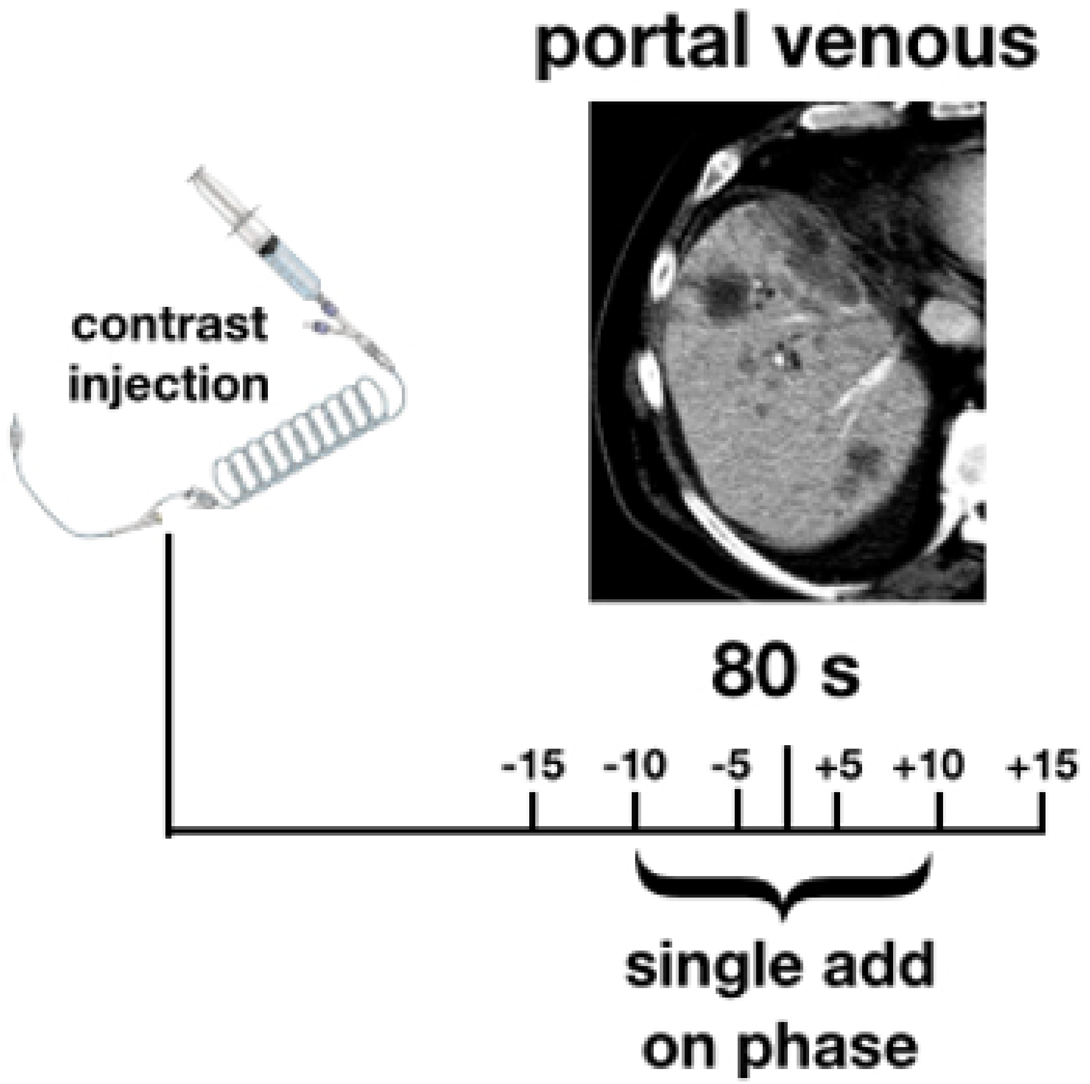
5 mm

0% ASiR (FBP)

60% ASiR

“Espresso Break” Reproducibility Study for Liver Parenchyma and Tumour Radiomics

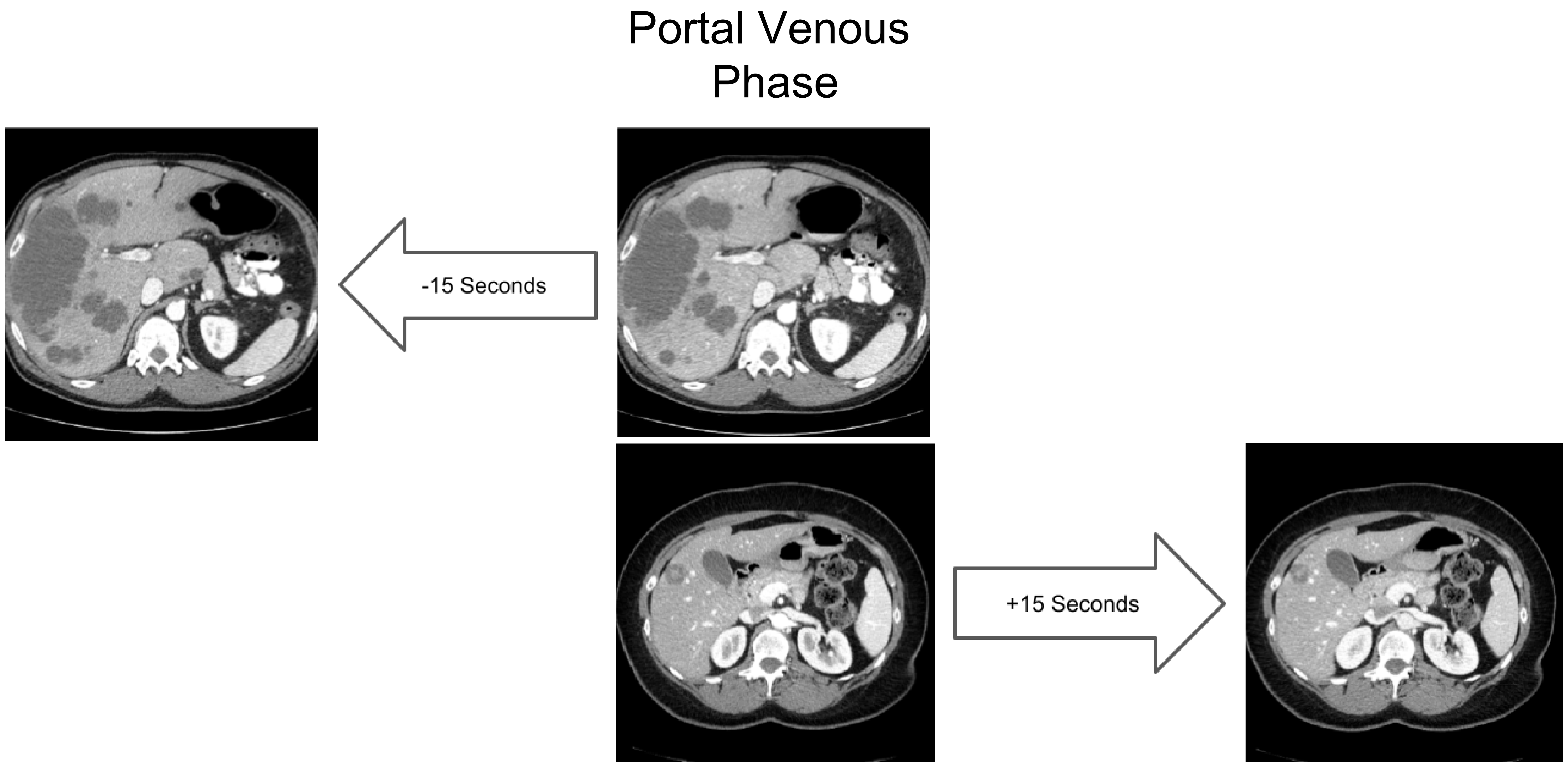




Do et al., RSNA 2024 and Virani-Wall et al. SPIE Medical Imaging 2025.

“Espresso Break” Reproducibility Study for Liver Parenchyma and Tumour Radiomics

- Contrast fluid is administered to patients prior to scan
- Liver attenuation is dynamic with respect to time



Do et al., RSNA 2024 and Virani-Wall et al. SPIE Medical Imaging 2025.

Sources of Statistical Bias & Variability – Data Acquisition

Data collection variability: Correction factors

- Example: Race-specific estimations of kidney function (EGFR)
- Fewer Black patients eligible for kidney transplant

Evaluating the Impact and Rationale of Race-Specific Estimations of Kidney Function: Estimations from U.S. NHANES, 2015-2018

What about the legacy data?

Download PDF Cite Share Set Alert Get Rights Reprints

Show Outline

ABSTRACT

Background

Standard equations for estimating glomerular filtration rate (eGFR) employ race multipliers, systematically inflating eGFR for Black patients. Such inflation is clinically significant because eGFR thresholds of 60, 30, and 20 ml/min/1.73m^2 guide kidney disease management. Racialized adjustment of eGFR in Black Americans may thereby affect their clinical care. In this study, we analyze and extrapolate national data to assess potential impacts of the eGFR race adjustment on qualification for kidney disease diagnosis, nephrologist referral, and transplantation listing.

Sources of Statistical Bias & Variability – Data Acquisition

Data collection variability: Race correction factors

The NEW ENGLAND JOURNAL of MEDICINE

CURRENT ISSUE SPECIALTIES TOPICS MULTIMEDIA LEARNING/CME AUTHOR CENTER PUBLICATIONS

MEDICINE AND SOCIETY

Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms

Authors: Darshali A. Vyas, M.D., Leo G. Eisenstein, M.D., and David S. Jones, M.D., Ph.D. Author Info & Affiliations

Published June 17, 2020 | N Engl J Med 2020;383:874-882 | DOI: 10.1056/NEJMms2004740 | VOL. 383 NO. 9

Sources of Statistical Bias & Variability – Data Acquisition

Data collection variability: Racial categories are incorrect

A second problem arises from the ways in which racial and ethnic categories are operationalized. Clinicians and medical researchers typically use the categories recommended by the Office of Management and Budget: five races and two ethnicities. But these categories are unreliable proxies for genetic differences and fail to capture the complexity of patients' racial and ethnic backgrounds.[34,35] Race correction therefore forces clinicians into absurdly reductionistic exercises. For example, should a physician use a double correction in the VBAC calculator for a pregnant person from the Dominican Republic who identifies as black and Hispanic? Should eGFR be race-adjusted for a patient with a white mother and a black father? Guidelines are silent on such issues — an indication of their inadequacy.

Sources of Statistical Bias & Variability – Preprocessing

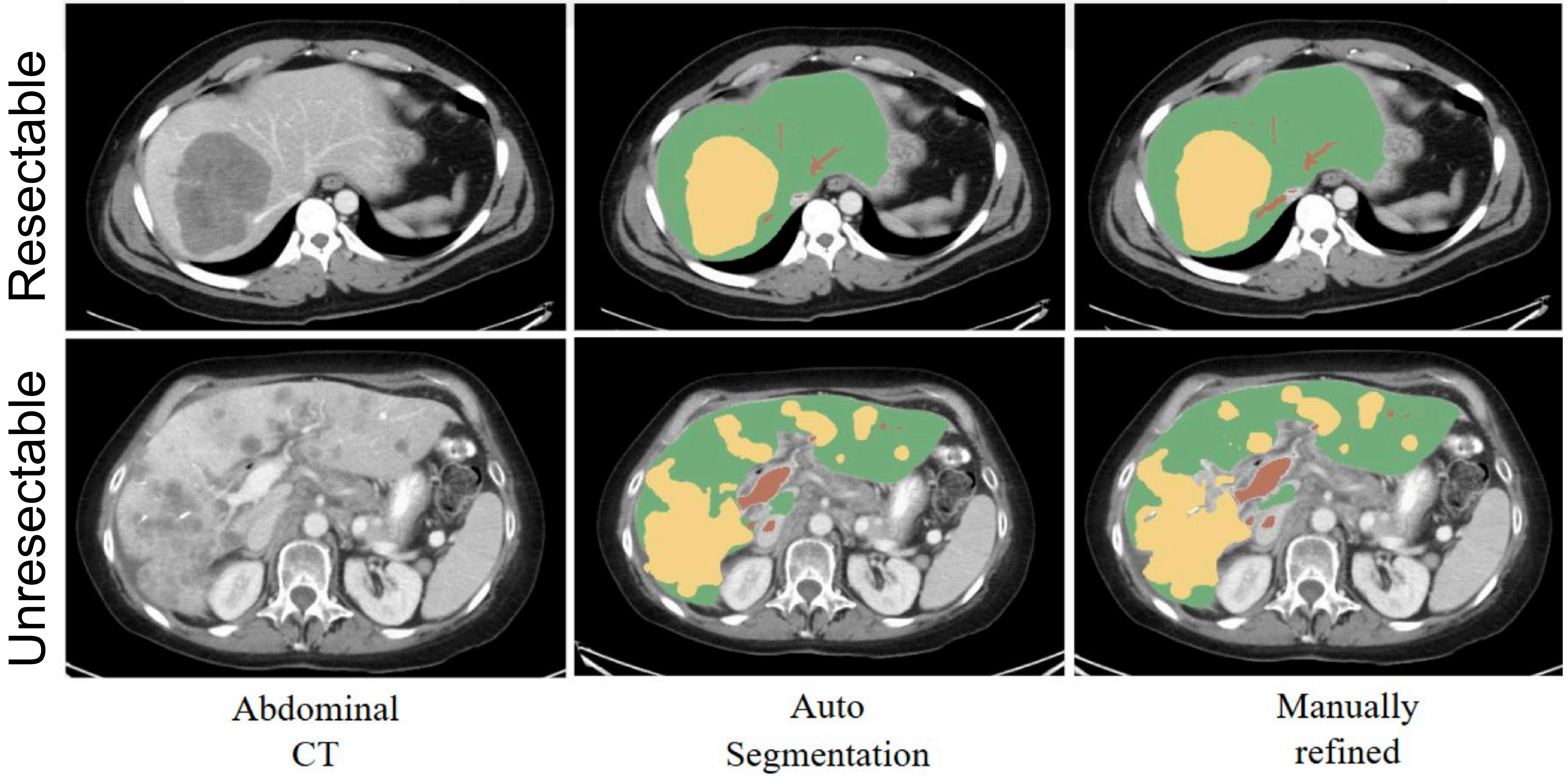
Analysis variability: Variability that arises when different filters, thresholding, etc. give different results

- Example: Image features vary based on the discretization method (i.e. fixed bin width or fixed number of bins)
- Advice: Make sure your methods state how the acquisition and pre-processing is performed so others can replicate

Sources of Statistical Bias & Variability – Preprocessing

Operator variability: Manual or semi-automated measurements differ based on human factors

- Example: Variability in image segmentation; this variability is also influenced by the disease site and existing clinical contour guidelines

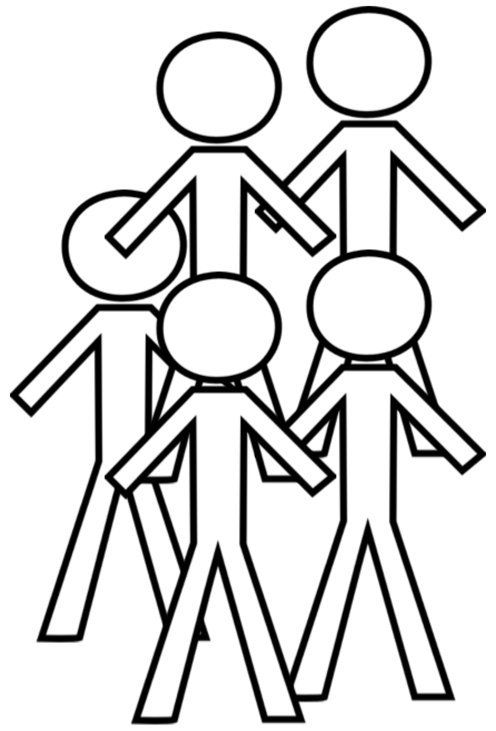


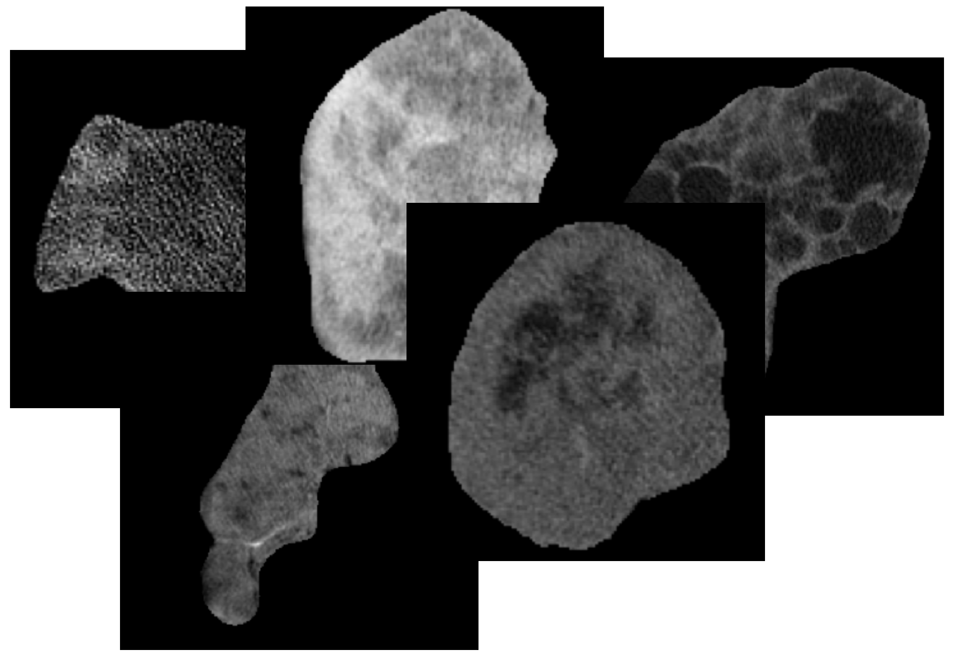
Sources of Statistical Bias & Variability - Statistical Analysis

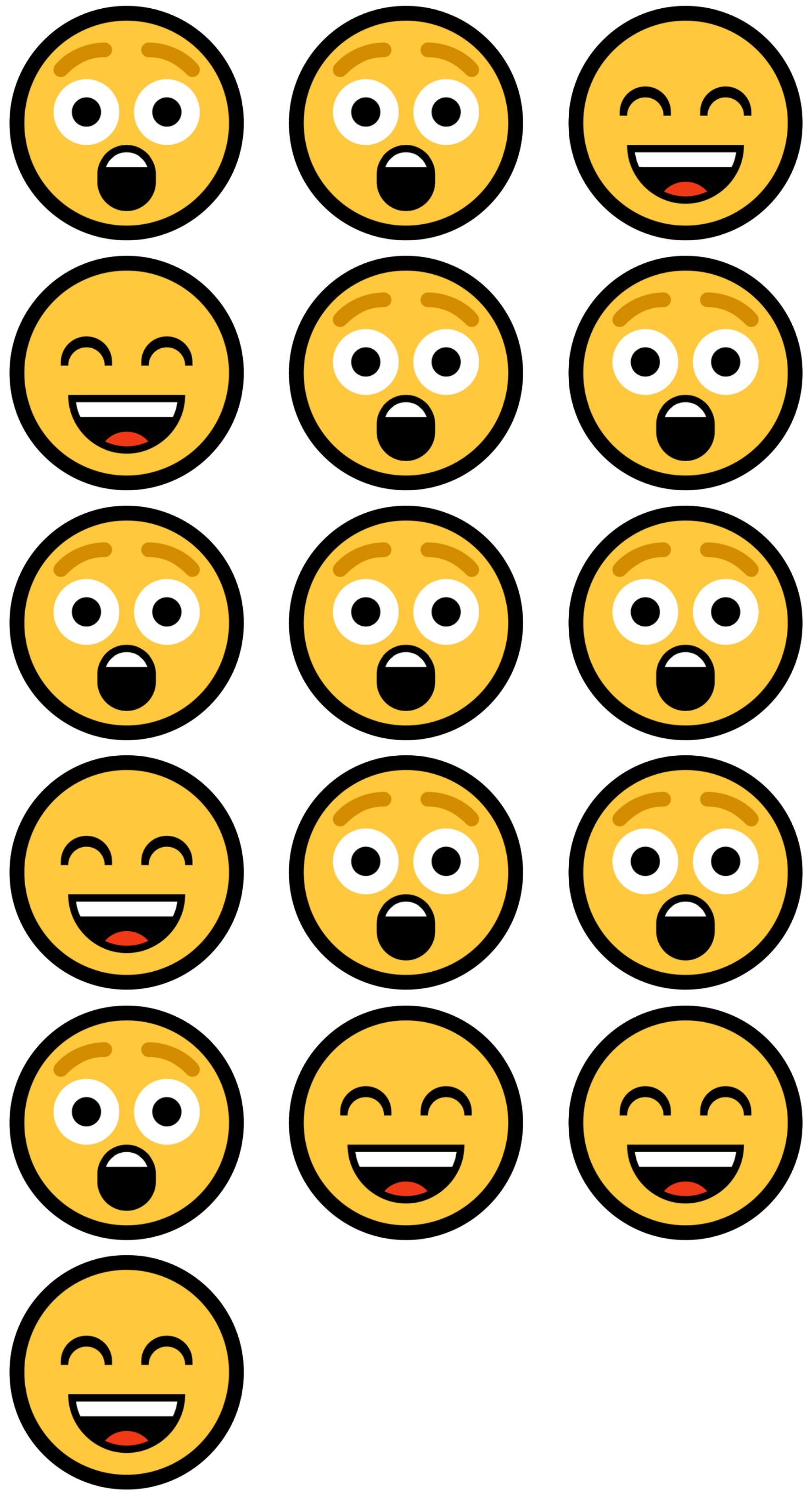
Bias due to overfitting: Model captures spurious associations in the training data, in addition to associations that would be replicated in similar datasets

Source: https://medium.com/analytics-vidhya/underfitting-vs-overfitting-vs-best-fitting-in-machine-learning-91bbabf576a5

Data Overfitting in Practice





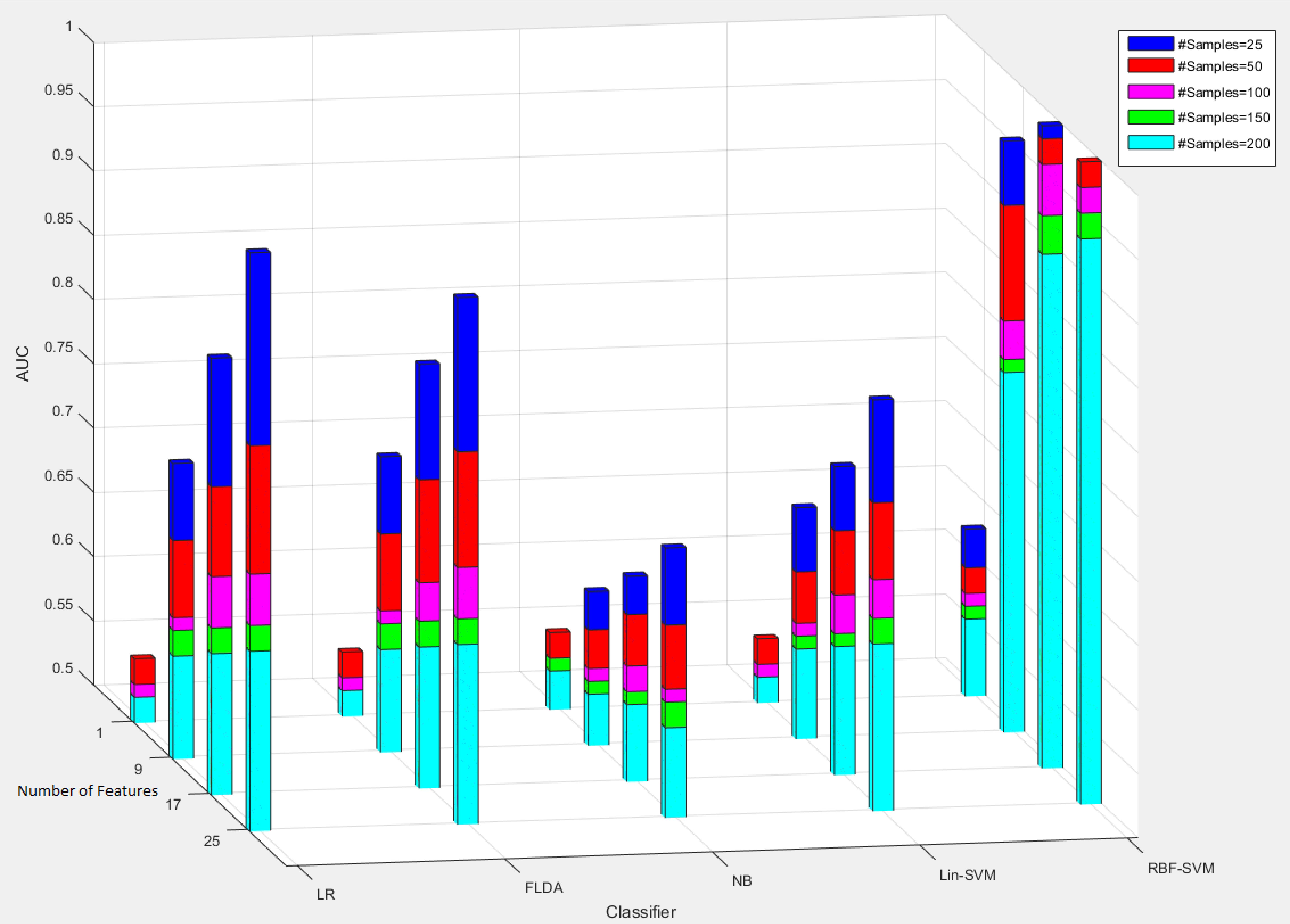


AUC = 0.5

Extract Tumor

Randomize Outcome

Build Prediction Model



We Can Predict Anything!

Chakraborty, Jayasree, et al. “Use of Response Permutation to Measure an Imaging Dataset’s Susceptibility to Overfitting by Selected Standard Analysis Pipelines.” *Academic Radiology* 31 (9): 3590–96, 2024

Sources of Statistical Bias & Variability

Statistical Analysis

- Optimistic performance bias: Evaluating the algorithm on the same data used to build or optimize the algorithm
    - Example: A model is developed to optimize performance in the training data. Model performance is assessed using both training the training and validation data.
    - Example: Feature selection performed on the full data set.

Data are incorrect and biased.

We already know a lot about statistical bias.

There is risk in doing nothing.

Decolonized AI Ethics

- AI bias supplement from NIH
- Digital Twin Podcast: https://www.queensu.ca/health-innovation/digital-cancer-twin-project/
- The Responsible Use of AI Podcast: https://podcast.cfrc.ca/podcasts/the-responsible-use-of-ai-podcast/
- Race correction + AI, indigenous data sovereignty


Annabelle Suave
MSc - CS


C. Stinson
AI Ethics


Vanessa Ferguson
MA - Phil


Jordan Loewen
Post Doc


Robyn Rowe
Post Doc


L. James
Health Data
Justice Postdoc

NIH NATIONAL CANCER INSTITUTE

SSHRC CRSH


S. Mosurinjohn
Humanist
School of Religion

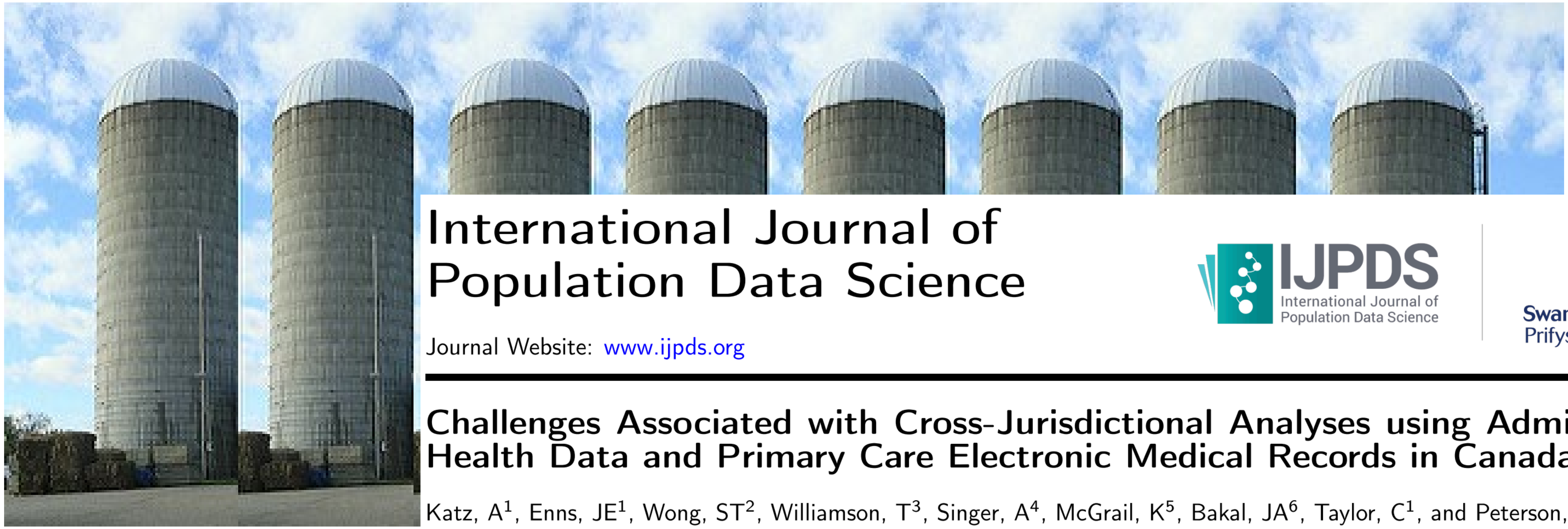

Ontario
Health • Santé
SIMPSON, AMBER L

© 2000. Her Majesty the Queen in Right of Canada, Natural Resources Canada.
Sa Majesté la Reine du chef du Canada, Ressources naturelles Canada.

Single payer insurance -> OHIP
Hospital -> not for profit private corporations

Slide courtesy of Dr. Alice Wei, UHN

Silofication of Canada’s Health Data



International Journal of Population Data Science

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Journal Website: www.ijpds.org

Challenges Associated with Cross-Jurisdictional Analyses using Administrative Health Data and Primary Care Electronic Medical Records in Canada

Katz, A[1], Enns, JE[1], Wong, ST[2], Williamson, T[3], Singer, A[4], McGrail, K[5], Bakal, JA[6], Taylor, C[1], and Peterson, S[2]

Submission History

Submitted:	30/11/2017
Accepted:	16/07/2018
Published:	05/10/2018

[1]Manitoba Centre for Health Policy, Department of Community Health Sciences, University of

Abstract

Over the last 30 years, public investments in Canada and many other countries have created clinical and administrative health data repositories to support research on health and social services, population health and health policy. However, there is limited capacity to share and use data across jurisdictional boundaries, in part because of inefficient and cumbersome procedures to access these data and gain approval for their use in research. A lack of harmonization among variables and

Our Lab


Claire Bunker
Undergrad - DBMS


Taryn Keenan
Undergrad - DBMS


John Zhou
Undergrad - CS


Mahmoud Idlbi
MSc - CS


Mane Piliposyan
MSc - CS


Shaina Smith
MSc DBMS


Dashti Ali
PhD - CS


Alan Dimitriev
PhD - CS


Andrew Garven
PhD - DBMS


Rina Khan
PhD - CS


Jaxen Smith
MSc - DBMS

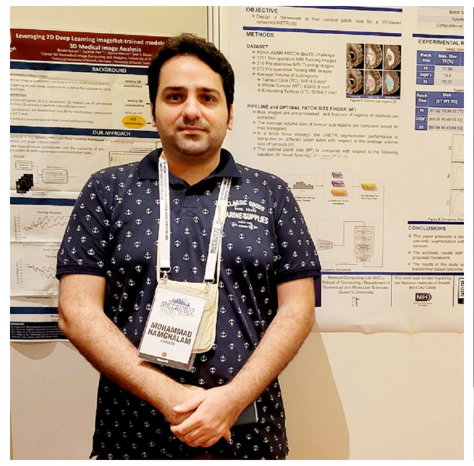

Josh Virani-Wall
MSc - CS


Kaitlyn Kobayashi
PhD - DBMS


Ramtin Mojtahedi Saffari
PhD - CS


Jianwei Yue
PhD - CS


Alex Robins
PhD - PHS


Mohammad Hamghalam
Post Doc


Jacob Peoples
AI Scientist

Alumni

INNOVATION
Canada Foundation for Innovation
Fondation canadienne pour l'innovation

THE TERRY FOX RESEARCH INSTITUTE
L'INSTITUT DE RECHERCHE TERRY FOX

AACR American Association for Cancer Research

Ontario

MARATHON OF HOPE CANCER CENTRES NETWORK

SSHRC CRSH

NSERC CRSNG

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

CANADA RESEARCH CHAIRS
CHAIRES DE RECHERCHE DU CANADA

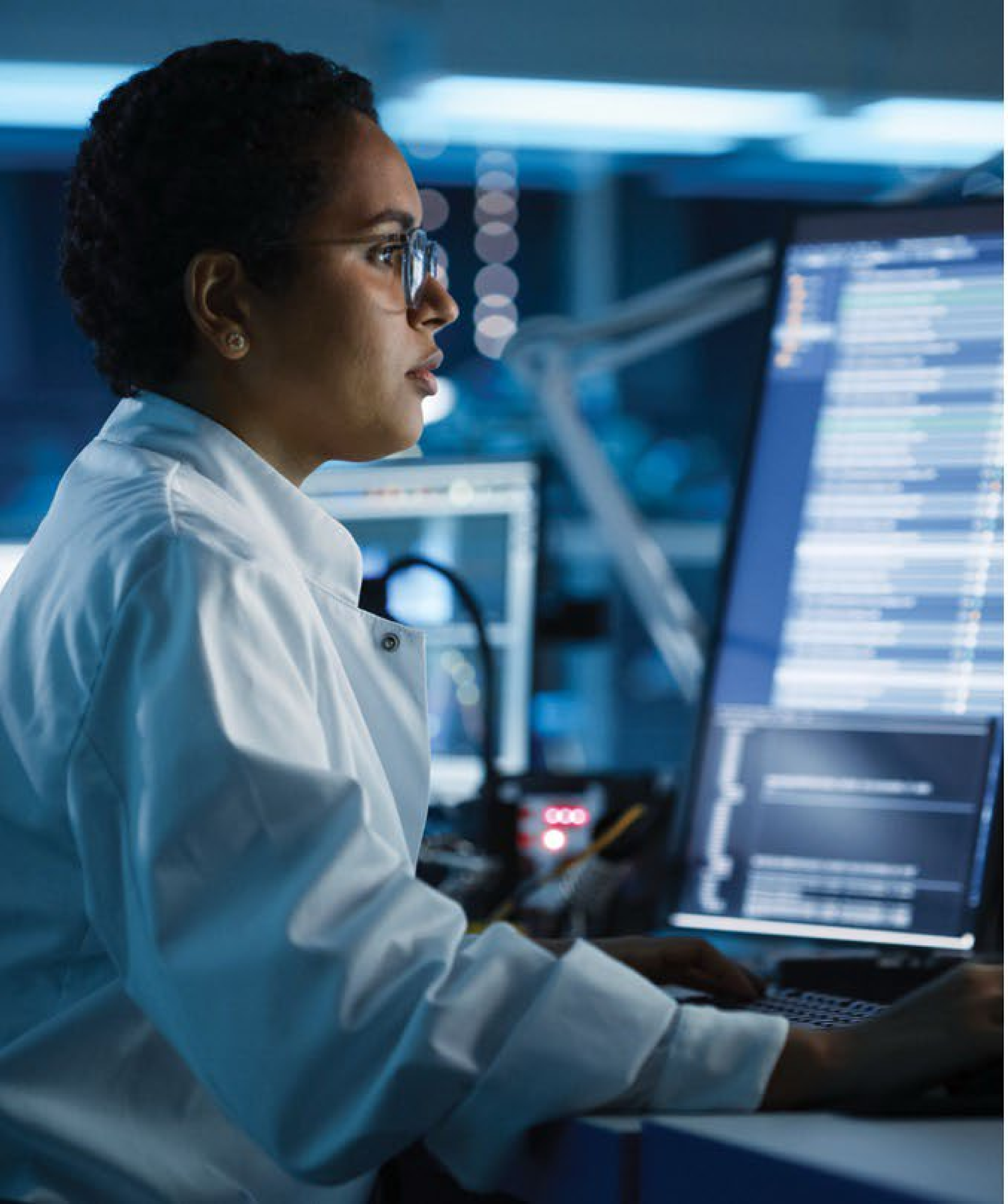
PANCREATIC CANCER ACTION NETWORK

SAR Society of Abdominal Radiology

CYCLE FOR SURVIVAL
EQUINOX

NIH NATIONAL CANCER INSTITUTE

How are the algorithms programming us?



CANCER RESEARCH

Special Series Call for Papers

Driving Cancer Discoveries with Computational Research, Data Science, and Machine Learning/AI

Scan to learn more.

AACR American Association for Cancer Research®

AACRJournals.org