CHALLENGES
IN
RADIOIMICS AND BIG DATA

Karen Drukker

ACKNOWLEDGEMENTS
• Lorenzo Pesce
• Maryellen Giger

FUNDING
• Supported over the years by The University of Chicago Dean Bridge Fund, grants from the NIH (NIH, NBR, NIAIMS), the DOD, the DOE, and the Segal Family Foundation, the UC Comprehensive Cancer Center, the SPORE, and the ITM/CTSA

CONFLICT OF INTEREST STATEMENT OF PRESENTER
• Karen Drukker receives royalties from Hologic

PURPOSE
• To summarize some of the statistical challenges in radiomics, genomics, radiogenomics, and big data from the perspective of a novice

† In this talk radiogenomics refers to the combination of radiomics features and genomic data, different from the use of this term in radiation oncology.
INTRODUCTION

WHAT IS RADIOMICS?

• Asks questions about the relationships between features "seen" in medical images and the biology of cancer

DEFINITIONS

• Radiomics: High throughput conversion of images to mineable data
  This data can be viewed as descriptors (i.e., phenotypes) of tumors and "normal" tissue

• Big Data: The exponential growth in the numbers of patients and the data elements being harvested from each is known colloquially as "big data"
**INTRODUCTION**

**DEFINITIONS**

**Phenotype:**
- The observable physical or biochemical characteristics of an organism, as determined by both genetic makeup and environmental influences.
- The expression of a specific trait, such as stature or blood type, based on genetic and environmental influences.

---

**Medical image data**

**Histopathology, molecular classification**

**Radiologist**
- Qualitative assessment

**Computer**
- Radiomics, quantitative assessment

**Associations and/or classification relevant to clinical or biological questions**
- Develop predictive models, personalized medicine

---

**Which are correlated and which are synergistic?**

---

**Medical image data**

**Histopathology, molecular classification**

**Genomics**

**Radiologist**
- Qualitative assessment

**Computer**
- Radiomics, quantitative assessment

**Associations and/or classification relevant to clinical or biological questions**
- Develop predictive models, personalized medicine

---

**Which are correlated and which are synergistic?**
INTRODUCTION
EXAMPLES OF IMAGE-BASED PHENOTYPES
Radiologist-assessed qualitative semantic phenotypes

Round-oval  Irregular

Radiologist - Computer
Qualitative assessment  Radiomics

Medical image data  Histopathology, molecular classification

Which are correlated and which are synergistic?

Associations and/or classification relevant to clinical or biological questions
- Develop predictive models, personalized medicine

INTRODUCTION
EXAMPLES OF IMAGE-BASED PHENOTYPES

In radiomics, we obtain quantitative computer-extracted image-based phenotypes (sometimes also referred to as agnostic phenotypes)

Sphericity: 0.80; 0.85  Irregularity: 0.65; 0.78
INTRODUCTION

EXAMPLES OF IMAGE-BASED PHENOTYPES

• Characterization of tumor and/or parenchyma texture
• Margin irregularity and sharpness
• CT: Lesion size (RECIST, volume)
• MRI: Kinetic characterization (uptake, washout)
• Nuclear medicine: SUV

“What’s in a name—
A rose by any other name would smell as sweet”
(Shakespeare)

But now on to the actual talk

PURPOSE

• To summarize some of the statistical challenges in radiomics, genomics, radiogenomics, and big data from the perspective of a novice
“...... must grasp basic statistics – or sloppy science will continue to grow”

INTRODUCTION

CENTRAL QUESTION

• Most scientist have been taught some statistics along the way
  – But often the type of statistics is not relevant to their current work
  – And once in the lab, people generally just do what everyone else does, without necessarily understanding why

INTRODUCTION

CHALLENGES IN RADIOMICS

• Reproducibility
• “Big Data”
• Data sharing
• Standardization/harmonization

INTRODUCTION

CHALLENGES IN RADIOMICS

• Reproducibility
• “Big Data”
• Data sharing
• Standardization/harmonization

What is the extent of the problem?

A 2012 study of 53 landmark papers in basic cancer research was able to replicate the original results of just 6 of these studies


INTRODUCTION

REPRODUCIBILITY

What is the extent of the problem?

A 2012 study of 53 landmark papers in basic cancer research was able to replicate the original results of just 6 of these studies


INTRODUCTION

REPRODUCIBILITY

What is the extent of the problem?

A 2012 study of 53 landmark papers in basic cancer research was able to replicate the original results of just 6 of these studies

**INTRODUCTION**

**BIG DATA**

What are the challenges?

- Large data sets
  - Not from carefully controlled experiments
  - Non-uniformity in equipment manufacturer, imaging protocol, imaged population etc

**OUTLINE OF THE REST OF THIS PRESENTATION**

- Hypothesis testing
- More on correlation and causality
- Reproducibility — “Big Data”
  - Radiomics
  - Genomics
  - Radiogenomics
- Discussion

**HYPOTHESIS TESTING**
HYPOTHESIS TESTING

- Hypothesis testing is like a criminal trial: A defendant is considered not guilty until proven otherwise.
  
  **Null-hypothesis \( H_0 \)**
  
<table>
<thead>
<tr>
<th>Accept null hypothesis</th>
<th>( H_0 ) is true</th>
<th>Right decision</th>
<th>( H_0 ) is true</th>
<th>( H_0 ) is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquittal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right decision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right decision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I error</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong decision</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q: Why do so many colleges and grad schools teach \( p=0.05 \)?
A: Because that’s still what the scientific community and journal editors use.

Q: Why do so many people still use \( p=0.05 \)?
A: Because that’s what they were taught in college or grad school.

**The American Statistician, 70, 128-133, 2016**
Absence of evidence is not evidence of absence
Altman DG and Bland JM, British medical journal, 311, 485, 1995

The statistical crisis in science
Geldman A and Loken E, American Scientist, 2014

The fallacy of the null-hypothesis significance test
Rosnow RL, Psychological Bulletin, 97, 408-429, 1980

False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as statistically significant
Simmons JP et al, Psychological Science, 22, 1359-1366, 2011

Why most published research findings are false
Ioannidis JP, PlosOne, 2, e124, 2005

HYPOTHESIS TESTING

AS A Statement on Statistical Significance and P-Values

1. p-values can indicate how incompatible the data are with a specified statistical model
2. A p-value does not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
3. Scientific conclusions and business decisions should not be based on whether a p-value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

HYPOTHESIS TESTING

AS A Statement on Statistical Significance and P-Values

1. p-values can indicate how incompatible the data are with a specified statistical model
2. A p-value does not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
3. Scientific conclusions and business decisions should not be based on whether a p-value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis
**HYPOTHESIS TESTING**

**MULTIPLE HYPOTHESIS TESTING**

- Often a variety of tests for a variety of possible effects are applied to a single data set and only those yielding a significant result are reported. **WRONG**

- Cherry-picking promising findings, Data dredging, Selective inference, Significance shopping, P-hacking

One needs multiplicity correction procedures that control the family wise error rate or the false discovery rate.

---

**HYPOTHESIS TESTING**

**MULTIPLE HYPOTHESIS TESTING**

- Multiplicity correction procedures
  - "Traditional" methods to control the family-wise error rate such as Bonferroni-Holm seen by many as too conservative
  - The test fails to reject the null hypothesis (that there is no difference) when there really is a difference. **Wrong**
  - Methods to control the false discovery rate (e.g., the Benjamini-Hochberg procedure) may be more useful, but perhaps too lax?
  - The null-hypothesis is rejected more frequently, but at the cost of indicating a statistically significant difference when there is none. **Wrong**

---

**MORE ON CORRELATION AND CAUSALITY**
MORE ON CORRELATION AND CAUSALITY

Association

Correlation

Causation


points of significance
Association, correlation and causation

Trends (linear or increasing/decreasing)

Uncorrelated

Correlated

Dependence

Associated

Not associated

Correlation

Causation

Null hypothesis: 2 methods/variables are not linearly related

A high correlation does not necessarily mean agreement

But even if the points lie along the equality line

Is this agreement "good enough"?

More insightful way of assessing agreement

Whether this is "good enough" depends on clinical task

Expected range in differences between measurements

More on Correlation and Causality
**More on Correlation and Causality**

Very different datasets may have similar/identical correlation coefficients.

Correlation also depends on:
- Noise
- Sample size

---

**More on Correlation and Causality**

Spurious correlation coefficients in random (independent) data.

Distribution of correlation coefficients.

95% confidence intervals of correlation coefficients.

---

**More on Correlation and Causality**

Why is this important to us?
- When the number of features is large with respect to the sample size, large but spurious correlations frequently occur.
- When there is a large number of observations, small and substantively unimportant correlations may appear statistically significant.

- "Small data" pilot studies
- "Big data" standard of care

---
**Radiomics**

**Helpful literature on statistics and related topics**

- Multi-analysis of the technical performance of imaging procedures: Accuracy and interventional methodology
- The emerging fields of applications in radiomic biomarker development and validation for diagnostic and therapeutic applications
- Radiomics: From radiographic images to quantitative biomarkers
- Radiomics: A bridge between imaging and molecular biology
- Radiomics: From marker to score

Quantitative Imaging Biomarkers Alliance

---

**Challenge: 'Man versus Machine'**

In radiomics part of the gold-standard 'truth' is often subjective, i.e., based on human assessment:

- Where is the lesion in the image?
- What is part of the lesion and what is not?

Example: Size measurement

- Disagreement between ‘man’ and ‘machine’
- Inter-radiologist (and intra-radiologist) variability

**Challenge: ‘Small data’ vs. ‘Big data’**

<table>
<thead>
<tr>
<th>Big data</th>
<th>Medium data</th>
<th>Small data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard of care</td>
<td>Clinical trials</td>
<td>Pilot studies</td>
</tr>
<tr>
<td>'Real world'</td>
<td>Carefully controlled</td>
<td></td>
</tr>
<tr>
<td>Large and very</td>
<td>experiment</td>
<td>Will my results be</td>
</tr>
<tr>
<td>diverse datasets</td>
<td></td>
<td>reproducible and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>generalizable?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inclusion and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>exclusion criteria</td>
</tr>
</tbody>
</table>

If the sample does not accurately represent the population of interest, statistics are not meaningful.
Many quantitative computer-extracted image-based features usually describe a single physical characteristic.

- The "machine" is able to extract much more, and perhaps more useful, information than the "human" (hundreds or even thousands of quantitative image-based features, i.e., phenotypes).
- In practice, often data sets of limited size are available for research and one needs to be able to obtain a realistic estimate of performance in the "real world".

Ideally use large training, calibration, and independent test data sets.

- In practice, modest-sized data sets using cross-validation or bootstrapping can give reasonable performance estimates when used properly.
- Reduce the number of computer-extracted features through supervised feature selection or unsupervised dimension reduction/clustering.
  - Parametric stochastic neighborhood embedding, Laplacian eigenmaps...
  - Use "smart" classifiers
    - Bayesian neural network, support vector machine, random forest...

Unsupervised dimension reduction (Laplacian eigenmaps)
- 1000+ lesions
- 81 image-based features

**RADIOMICS**

**‘BIG DATA’ CHALLENGE: HARMONIZATION AND INTERPRETATION OF RESULTS**

Differences in image acquisition or the population may affect computer-extracted image-based features

- Manufacturer
- Imaging protocol
- Geographic location
- Racial differences in disease prevalence and characteristics
- Actual outcome data such as survival may not be available and intermediate alternatives may need to be used
- Need harmonization, especially for less standardized modalities such as MRI
- Correlations amongst the many observations may be vast in number but spurious or unimportant, causality is much harder to assess

N Gruszczak et al, Radiology, 253, 661-671, 2009

---

**RADIOMICS**

**REPRODUCIBILITY INVESTIGATIONS**

Yesterday morning’s lecture by Laurence Court

---

**RADIOMICS**

**REPRODUCIBILITY INVESTIGATIONS**

A. Chalkidou et al, PLoSONE 10(5): e0124165

---
**Radiomics Reproducibility Investigations**

- Example of a relatively large radiomics study
  - Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach
  - Aerts HJWL et al., Nature Communications, DOI: 10.1038/ncomms5006, 2014
Radiomics

Example of a relatively large radiomics study

Take home message:
- Modest but robust performance for 2 different cancer types in independent datasets
- So... if the results of other studies sound too good to be true, they probably are?

Radiomics

Example of a relatively large radiomics study

Reproducibility
Genomics
Challenges here are amplified with respect to radiomics

- Whole exomes (20-30K genes), whole genomes (3 billion base pairs), with tens of millions of single nucleotide polymorphisms (SNPs) etc.
- Number of samples ~10⁶-10⁷nds

Need correction for multiple-hypotheses testing

Want short, reproducible, predictive gene lists (PGLs)

Hypothesis: Genes most important and relevant for control of the malignancy also appear on the list of the most predictive genes

Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer

<table>
<thead>
<tr>
<th>Authors</th>
<th>Technology</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al</td>
<td>Affymetrix</td>
<td>286</td>
</tr>
<tr>
<td>van't Veer et al</td>
<td>Rosetta microarrays</td>
<td>96 and 295</td>
</tr>
</tbody>
</table>

Example: survival prediction in breast cancer
GENOMICS

CHALLENGE: REPLICATION OF RESULTS AND MULTIPLE HYPOTHESIS TESTING

• Example: survival prediction in breast cancer

<table>
<thead>
<tr>
<th>Authors</th>
<th>Technology</th>
<th>Number of patients</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al</td>
<td>Affymetrix</td>
<td>286</td>
<td>76</td>
</tr>
<tr>
<td>van’t Veer et al</td>
<td>Rosetta microarrays</td>
<td>96 and 295</td>
<td>70</td>
</tr>
</tbody>
</table>

• Example: survival prediction in breast cancer

Why these differences?!

• Can NOT be explained by only
  • Differences in patient cohorts (such as age)
  • Different microarrays used
  • Different methods of data analysis

Performance was not reproduced, i.e., substantially worse, when tested on different dataset

Only 3 genes in common!

Different PGLs obtained from different training sets generated from the same patient cohort typically only have a few genes in common

• Model for the overlap of top PGLs obtained in replication studies (based on “probably approximately correct” sorting)
• Expected overlap differs for different cancer types

Ein-Dor et al., PNAS, 103, 5923-28, 2006

GENOMICS
CHALLENGE: REPLICATION
OF RESULTS AND MULTIPLE HYPOTHESIS TESTING

• Common practice in large-scale genomic studies to use p-values to choose which of numerous hypothesis test results should be pursued in subsequent research

• But... p-values themselves are highly variable

P-values are data dependent statistics that vary from sample to sample even when underlying effects, population, and sampling are the same

GENOMICS
CHALLENGE: REPLICATION OF RESULTS AND MULTIPLE HYPOTHESIS TESTING

• The good news: P-values are not only highly variable, but the degree of variability is
  • predictable
  • consistent across most types of statistical studies
The prediction intervals for p-value variability can be computed using only
  • the p-value of the original study
  • the sample size of the replication study relative to that of the original study

Lazzeroni et al., Molecular Psychiatry, 19, 1336-1340, 2014

GENOMICS
CHALLENGE: REPLICATION...

• Alzheimer study
• 488,512 SNPs
• 939 individuals
• Original p-value ~10^-8 for a specific gene association with entorhinal cortical volume

P-value variability prediction for replication study of equal size as original study

SJ Furney et al., Mol Psychiatry 16, 1130-1138, 2001
Lazzeroni et al., Molecular Psychiatry, 19, 1336-40, 2014

GENOMICS
CHALLENGE: REPLICATION...

• Alzheimer study
• 488,512 SNPs
• 939 individuals
• Original p-value ~10^-8 for a specific gene association with entorhinal cortical volume

P-value variability prediction for replication study of different sizes than original study

SJ Furney et al., Mol Psychiatry 16, 1130-1138, 2001
Lazzeroni et al., Molecular Psychiatry, 19, 1336-40, 2014
**Genomics**

**Challenge: Replication**
- Alzheimer study
- 488,911 SNPs
- 939 individuals
- Original p-value ~10^{-8} for a specific gene association with entorhinal cortical volume

95% confidence intervals for most significant p-value reported in original study

Lazarro et al., Molecular Psychiatry, 19, 1336-40, 2014

---

**Reproducibility**

**Radiogenomics**

---

**Radiogenomics**

---
**Radiogenomics**

**Challenges: All challenges of radiomics and genomics combined...**

**Medical Imaging data**

The Cancer Imaging Atlas (TCIA)

**Molecular/genetic data**

Y. Yuan website

---

**Radiogenomics**

**Example exploratory study**

Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data

Radiomics: MR image-based phenotypes [38 features]

Genomics: genes from 2 recent breast cancer studies [144 features]


---

**Radiogenomics**

**Example exploratory study**

Small study: 91 invasive breast cancers from TCGA/TCIA

Study design limits overtraining / database bias through

- the inclusion of only those genes previously identified by others as potentially useful
- two-tier cross-validation

Investigate the relation between radiomics, genomics, and combined radiogenomics features and estrogen receptor, progesterone receptor, human epidermal growth factor, tumor stage, and lymph node status.

Significant associations between radiomic features and clinical outcomes were observed. Overall, the prediction performances of genomics alone, radiomics alone, and combined radiogenomics features showed statistically significant correlations with clinical outcomes.

DISCUSSION
DISCUSSION

So...?

Plenty of challenges remain, including computational and statistical aspects. We need

• Harmonization, standards
• Full-disclosure reporting: the way in which predictor variables were chosen and the description of applied significance tests should be as clear as the way in which a dataset was chosen
• Better access to raw data such as publicly available data sets to serve as independent benchmarks

Challenges fuel progress!

THE END

Thank you

RADIOMICS

REPRODUCIBILITY INVESTIGATIONS

A. Chalkidou et al. *Radiology* 10(5): e022405
**Genomics**

**Challenge:** Comprehensive prediction of cancer genomic interactions

**A solution:** Zodiac (www.compgenome.org/zodiac)

- Interactions inferred by Bayesian Graphical Model

---

**Imaging Genomics**

**Challenge:** Provide a basis for using non-invasive imaging techniques to indirectly assess molecular evolution of tumors and their changes under treatment

**Pilot study on breast tumors**

Quantitative analysis of TCIA breast MRI cases (Giger lab)

Analysis of genomics data from corresponding TCGA breast cases (Ji lab)

Imaging Genomics analysis of 91 TCGA breast cancer cases

---

**Imaging Genomics**

**Pathway transcriptional activities associated with MRI quantitative features**

---
IMPORTANT STATISTICAL CONCEPTS