When statistics meets bioinformatics

Patty Solomon
School of Mathematical Sciences

Undergraduate Seminar, 11 May 2011
The University of Adelaide

Lies, damn lies and statistics

Truths, damn truths and statistics

I. Innovation in statistics is best driven by substantive applications.

Illustration:

Sequencing of the human genome led to the design and analysis of microarray experiments

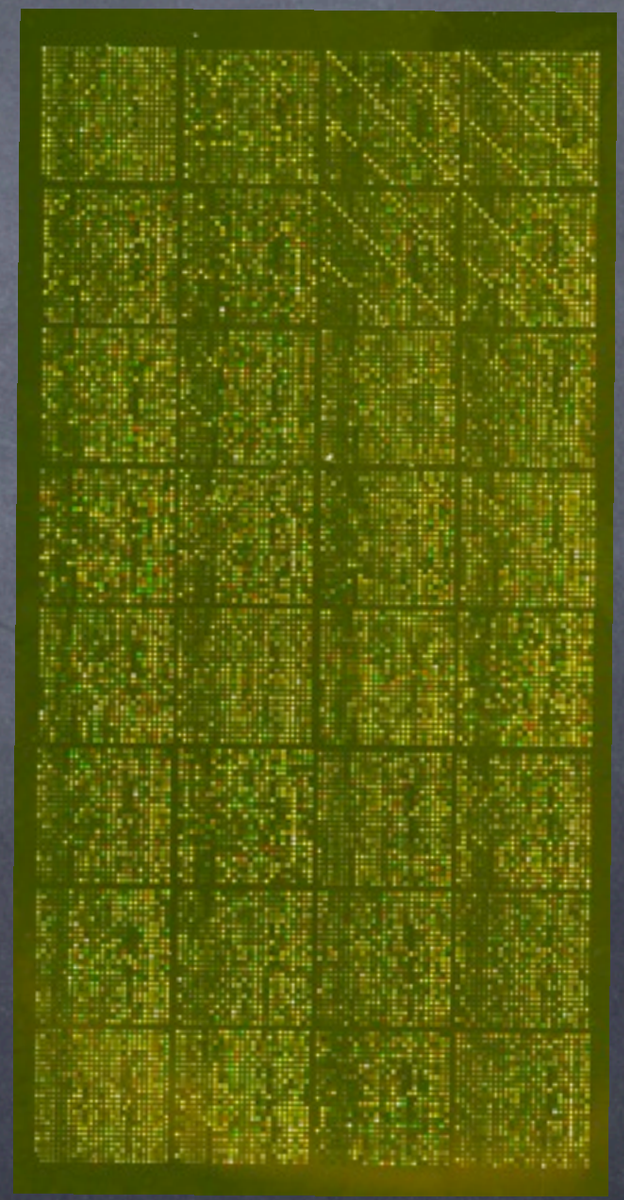
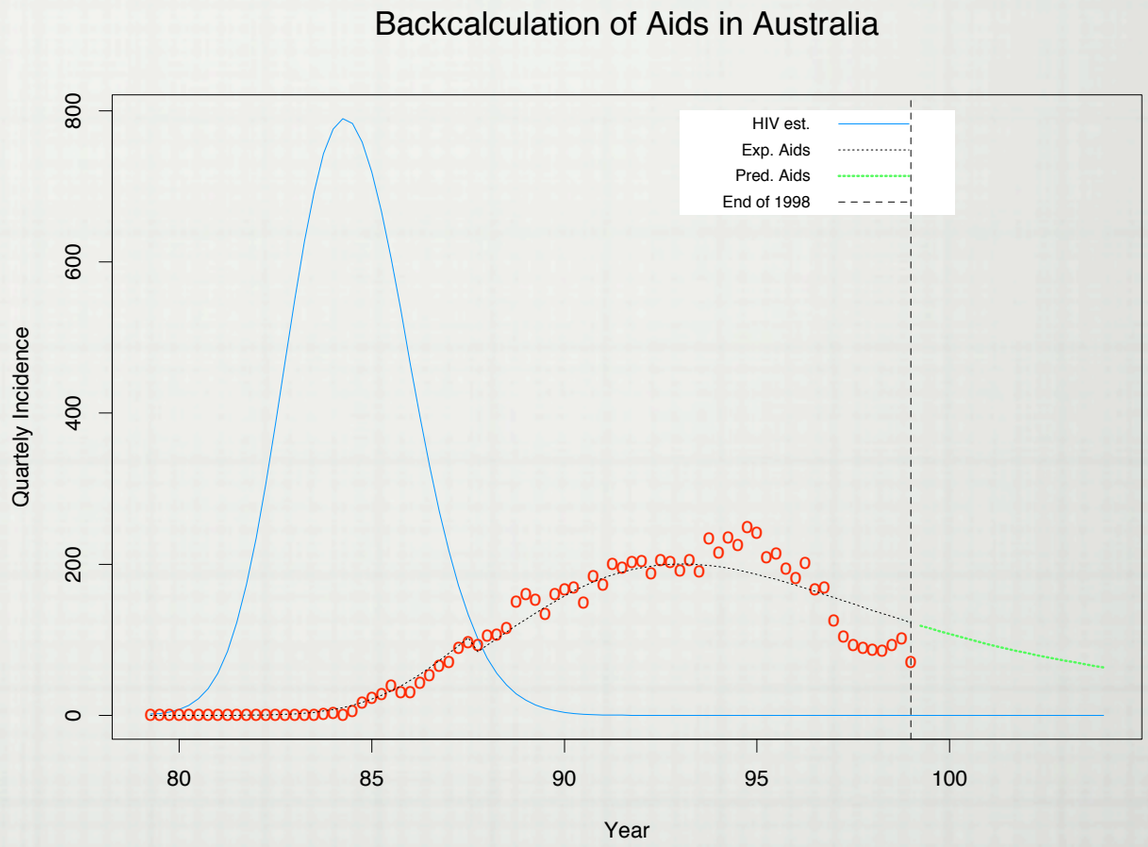


Image of a two-colour human microarray

INNOVATION IN STATISTICS CAN BE DRIVEN BY DISASTERS

- HIV/AIDS EPIDEMIC
- METHOD OF BACK-CALCULATION TO RECONSTRUCT HIV INFECTION INCIDENCE FROM OBSERVED AIDS INCIDENCE
- AS SUSCEPTIBLE-INFECTED-REMOVED EPIDEMIC MODEL:

$$a(t) = \int_0^t h(u) f(t-u) du$$



Some more truths

II. Biology is dominating statistics at the beginning of this century, just as it did at the beginning of the last one.

Why?

III. Statistics is an enabling discipline.

It has its own coherence as a mathematical discipline (like Pure Mathematics).
But good statistical analysis is the key to getting the best out of the new biotechnologies.
We have by training the skills of experimental design, data analysis, synthesis and reasoning which are essential to bioinformatics.

SOME BIOLOGICAL BACKGROUND

- ☐ DNA MAKES RNA MAKES PROTEIN.
- ☐ MICROARRAYS ALLOW US TO MEASURE THE MRNA COMPLEMENT OF A SET OF CELLS.
- ☐ MASS SPECTROMETRY ALLOWS US TO MEASURE THE PROTEIN COMPLEMENT OF A SET OF CELLS.
- ☐ PROTEOMIC SPECTRA ARE MASS SPECTROMETRY TRACES OF BIOLOGICAL SPECIMENS.

WHY ARE WE EXCITED?

- [ ] THERE IS A GREAT DEAL OF INTEREST IN DISCOVERING PROTEIN BIOMARKERS TO IDENTIFY CANCER PATIENTS EARLY ON.

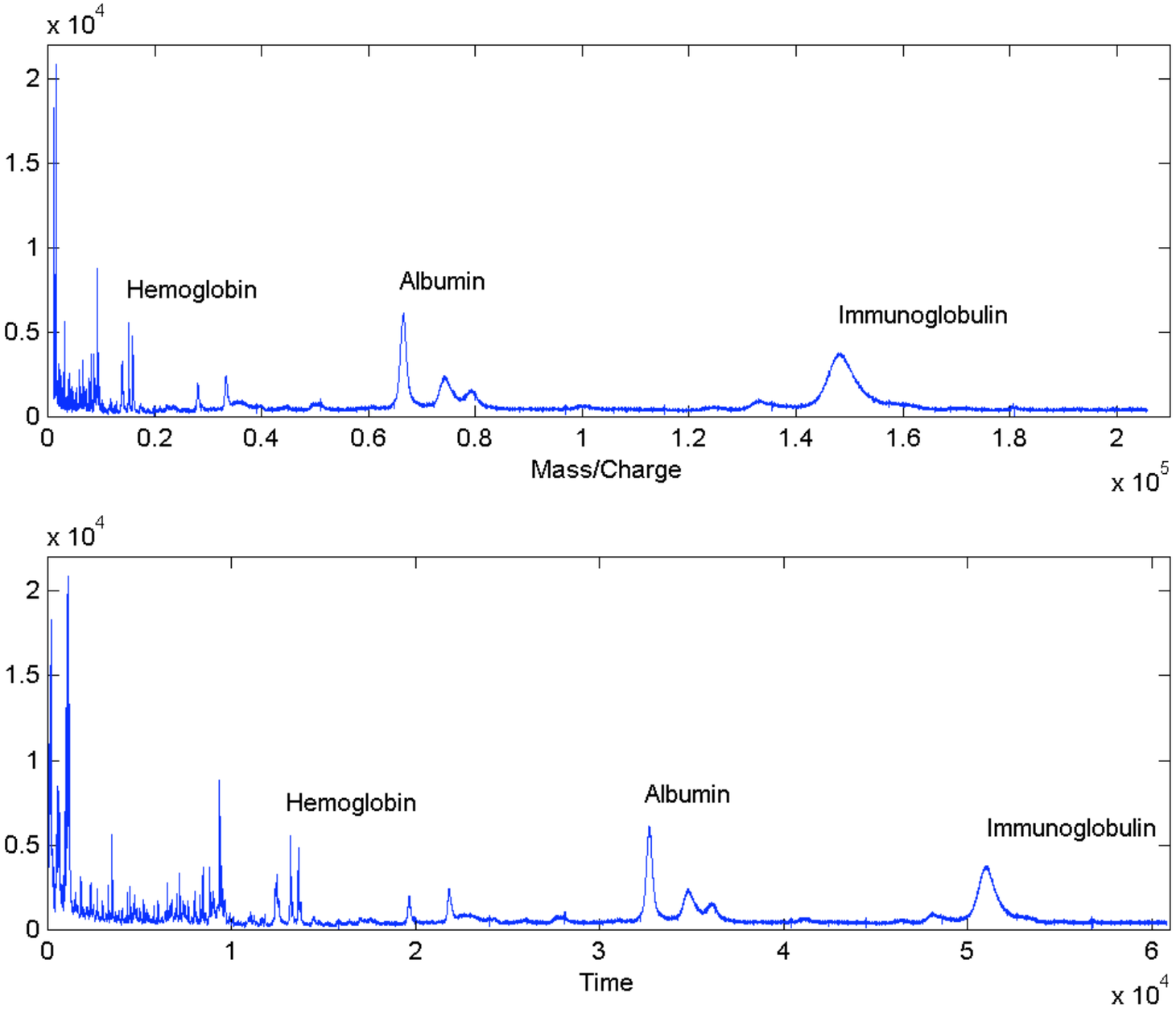
- [ ] PROTEIN PROFILES ARE BEING ASSESSED USING SERUM AND URINE, NOT TISSUE BIOPSIES.

- [ ] PROTEOMIC SPECTRA ARE CHEAPER TO RUN ON A PER UNIT BASIS THAN MICROARRAYS.

- [ ] CAN RUN SAMPLES ON LARGE NUMBERS OF PATIENTS.

WHAT DO THE DATA LOOK LIKE?

A MASS SPECTRUM OF HUMAN SERUM



OUR CASE STUDY

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 OVARIAN CANCER PATIENTS
- 100 NORMAL CONTROLS
- 16 PATIENTS WITH 'BENIGN' DISEASE.
- USED 50 CANCER AND 50 NORMAL SPECTRA TO TRAIN A CLASSIFICATION METHOD, AND THEN TESTED THE ALGORITHM ON THE REST OF THE DATA.

Model Selection

- The data were randomly split into a **training dataset** to fit the model, and a **test dataset** to estimate the Prediction Error (PE).

- This approach produces an unbiased estimate of the PE.

- In the methods you have met so far, e.g. linear regression, the training set = test set, which gives an overly optimistic estimate of the PE.

THEIR RESULTS

MECHANISMS OF DISEASE

Mechanisms of disease

Lancet, 359, 2002:572-7

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

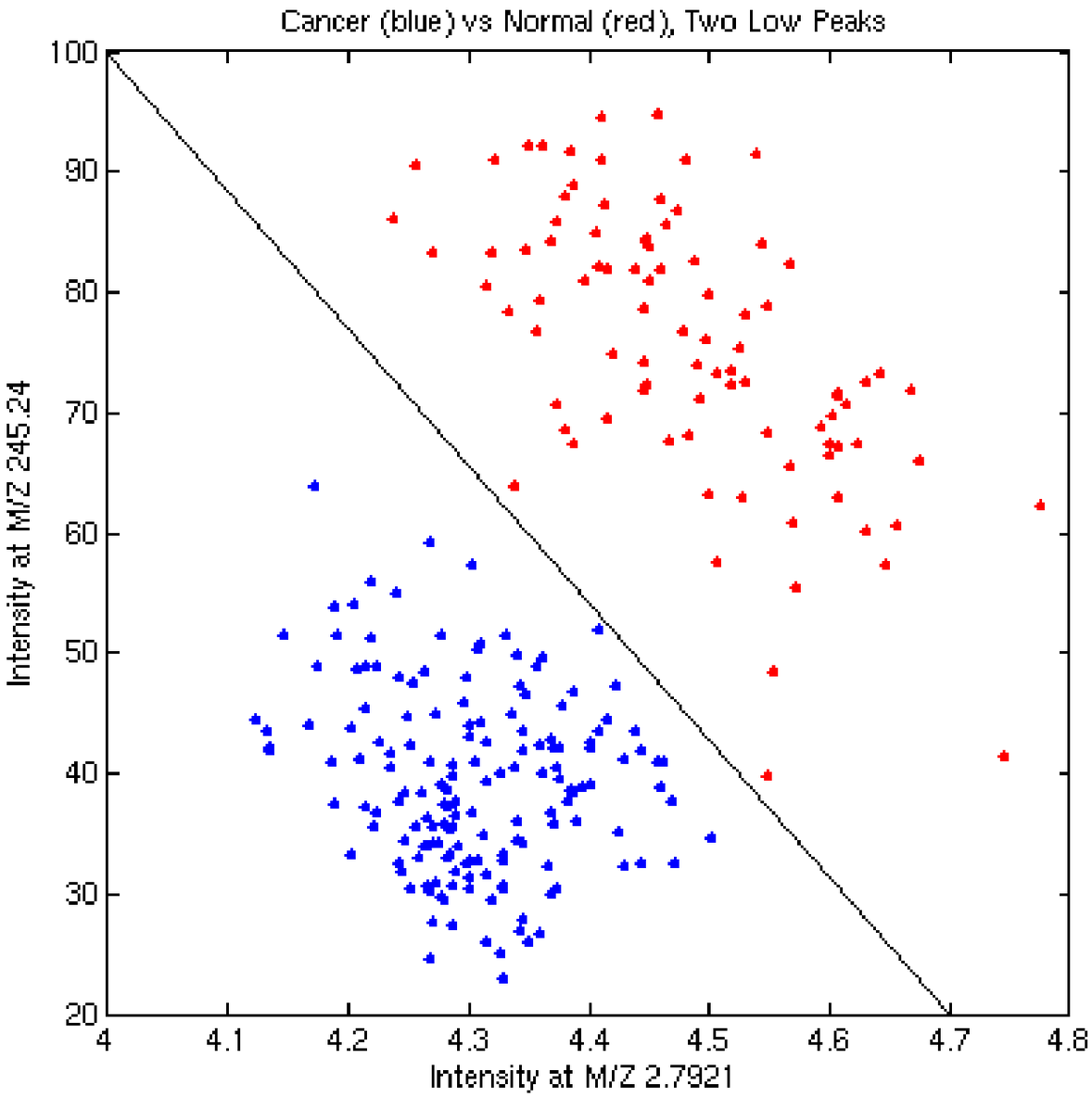
- CORRECTLY CLASSIFIED 50/50 OF THE OVARIAN CANCER CASES.
- CORRECTLY CLASSIFIED 46/50 OF THE NORMAL CASES.
- CORRECTLY CLASSIFIED 16/16 OF THE BENIGN DISEASE AS 'OTHER'.

MUCH EXCITEMENT ...

- ☐ GROUPS AROUND THE WORLD STARTED ASKING HOW TO DO THIS WITH THEIR TYPE OF CANCER ...
- ☐ BUT ALMOST IMMEDIATELY, EXPERTS RAISED OBJECTIONS ABOUT THE APPROACH: IT SHOULDN'T WORK, OWING TO LIMITATIONS OF THE TECHNOLOGY.
- ☐ VARIOUS QUESTIONS ABOUT ODDITIES IN THE DATA BEGIN TO CROP UP ...

- [ ] THE RESULTS WERE NOT REPRODUCIBLE FROM THE 'SAME' DATA.
- [ ] NO TIME-M/Z CALIBRATION.
- [ ] THERE WAS AN APPARENT CHANGE OF PROTOCOL NEAR THE END OF THE DATASET.
- [ ] NO EVIDENCE THAT THE ORDER OF PROCESSING WAS RANDOMISED.
- [ ] THERE IS NOTHING IN THE PAPER ABOUT THE SAMPLES, HOW THEY WERE COLLECTED OR PROCESSED, OR ANY CLINICAL OR DEMOGRAPHIC INFORMATION - ONLY THE PATIENT'S CASE/CONTROL STATUS IS REFERRED TO OR USED IN THE ANALYSIS.
- [ ] PERFECT CLASSIFICATION OF PEAKS IS ACHIEVED IN THE "NOISE" REGION OF THE DATA (SEE NEXT SLIDE ...)

Another Bivariate Plot: M/Z = (2.79,245.2)



Perfect Separation, using a completely different pair. Further, look at the masses: this is the noise region.

- ALL THIS (AND MORE) STRONGLY SUGGESTED A QUALITATIVE DIFFERENCE IN HOW THE SAMPLES WERE PROCESSED, AND POSSIBLY NOT JUST A DIFFERENCE IN THE BIOLOGY.

- IN JANUARY 2004 CORRELOGIC, QUESTDIAGNOSTICS AND LABCORP ANNOUNCED PLANS TO OFFER A 'HOME BREW' TEST CALLED OVACHECK: SAMPLES WOULD BE SENT BY CLINICIANS FOR DIAGNOSIS.

- ESTIMATED MARKET: 8 TO 10 MILLION WOMEN.

- ESTIMATED COST: US$100-200 PER TEST.

IN THE MEANTIME, AN ABORTIVE SECOND PAPER APPEARS …

- THE SAME GROUP* PROCESSED SAMPLES WITH THEIR ORIGINAL MS TECHNOLOGY AND ALSO WITH A HIGHER RESOLUTION INSTRUMENT (QQTOF). THEY ADDED SOME QUALITY CONTROL STEPS TO REMOVE BAD SPECTRA; STILL USING PATTERNS.

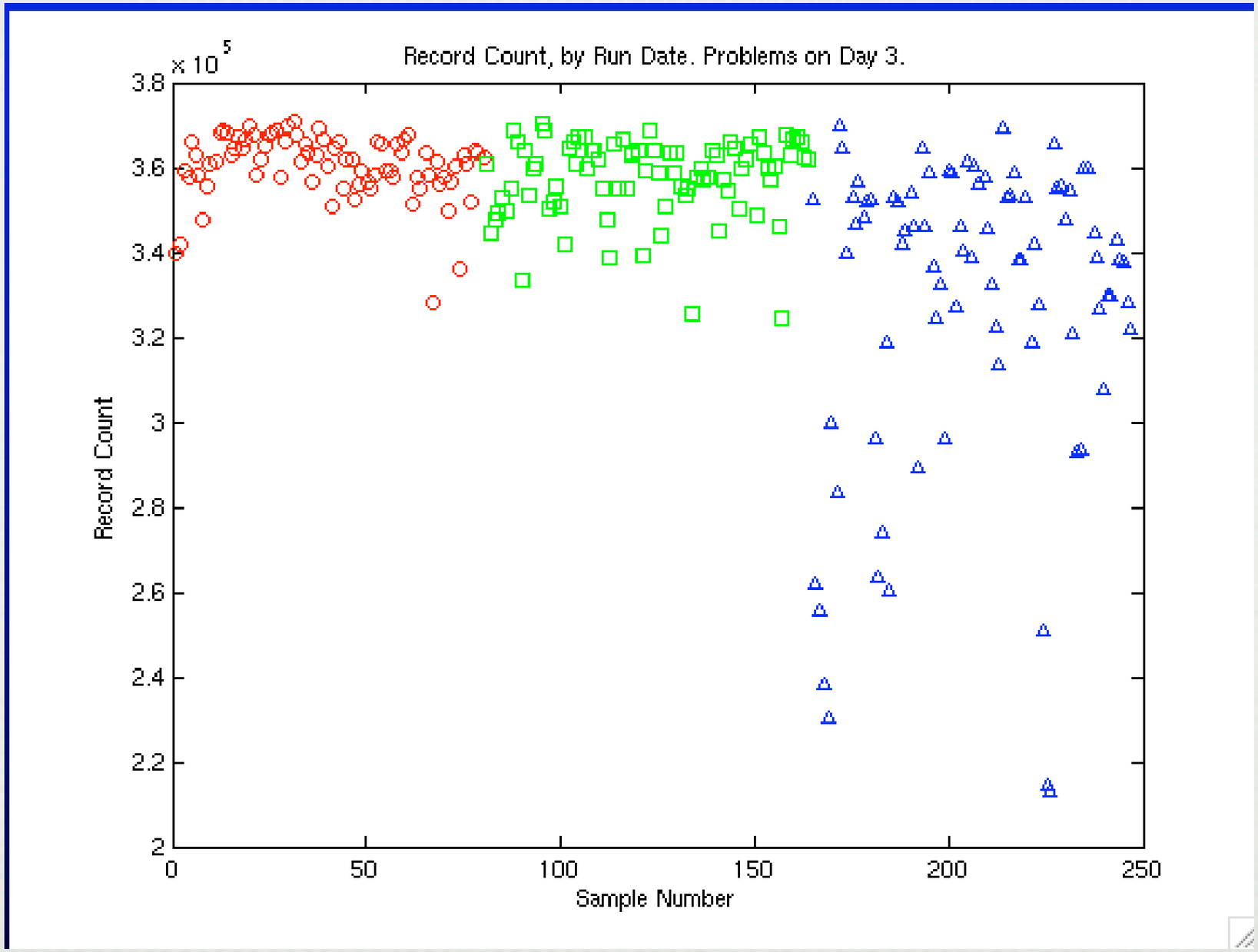
- THESE RESULTS WERE EVEN BETTER!

- 100% SENSITIVITY AND 100% SPECIFICITY FOR IDENTIFYING CANCER FROM NORMAL AND CLAIMED THIS "EMERGING PARADIGM" IS READY TO GO TO A LARGE CLINICAL STUDY.

SO WHAT WAS GOING ON?

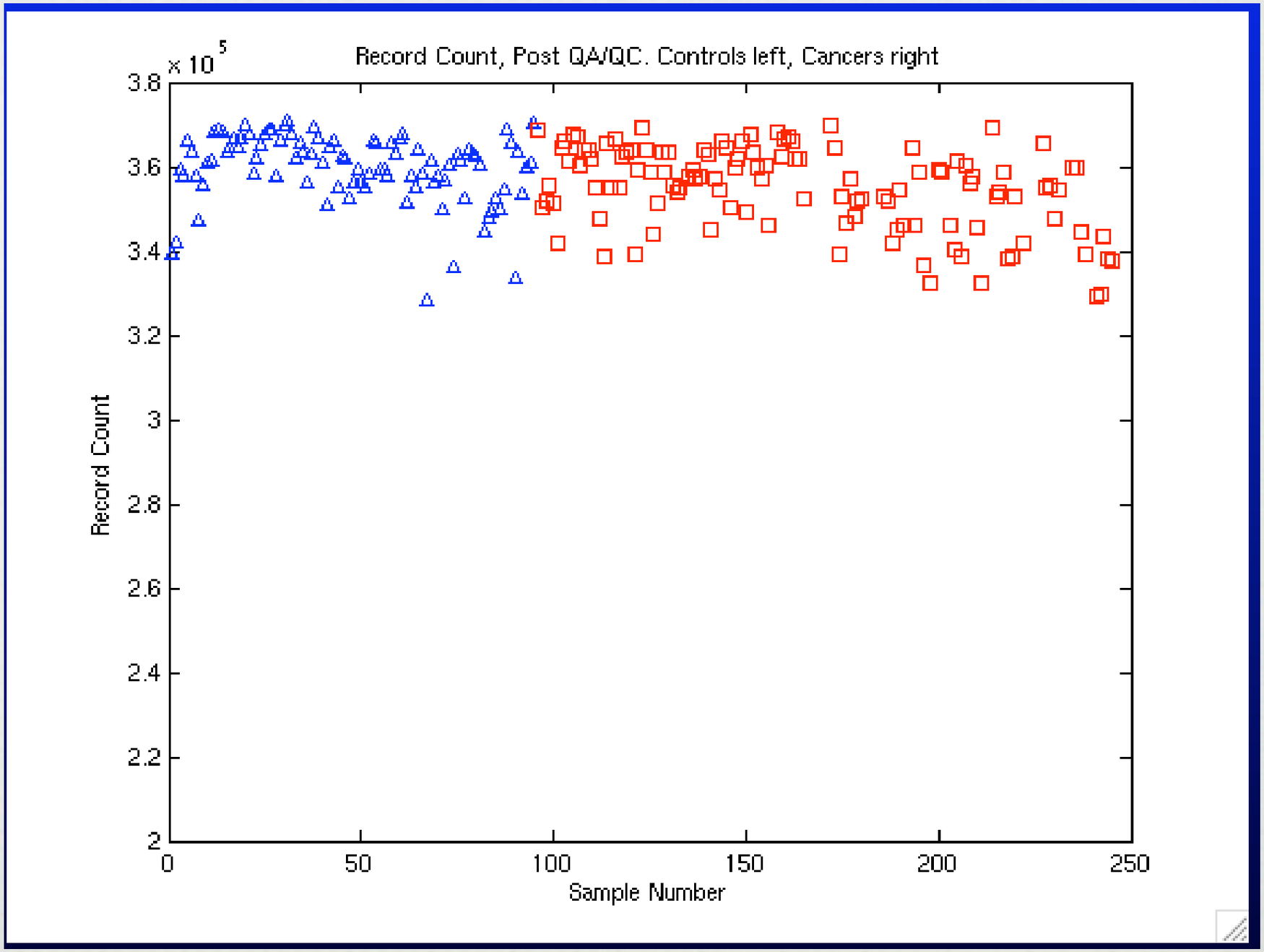
- * CONRADS ET AL, ENDOCRINE RELATED CANCER 11, 163-178, 2004

PART 1: HERE IS THEIR FIGURE 6A



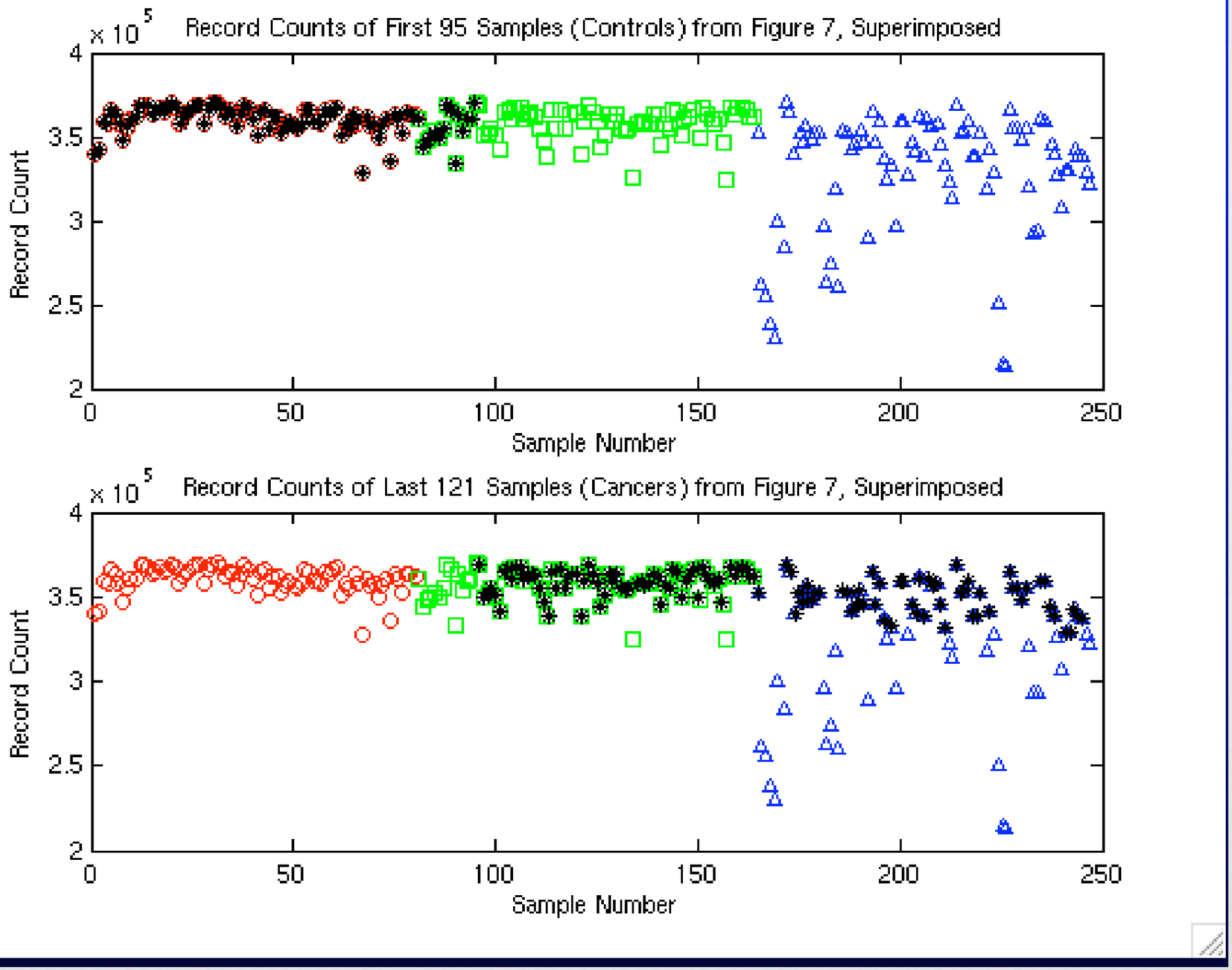
COLOUR = DAY 1, 2, 3

PART II: THEY DO SOMETHING ABOUT IT, FIGURE 7



(IN THE ORIGINAL PAPER, THE CASES AND CONTROLS ARE MISLABELLED)

PART III: WHAT'S GOING ON

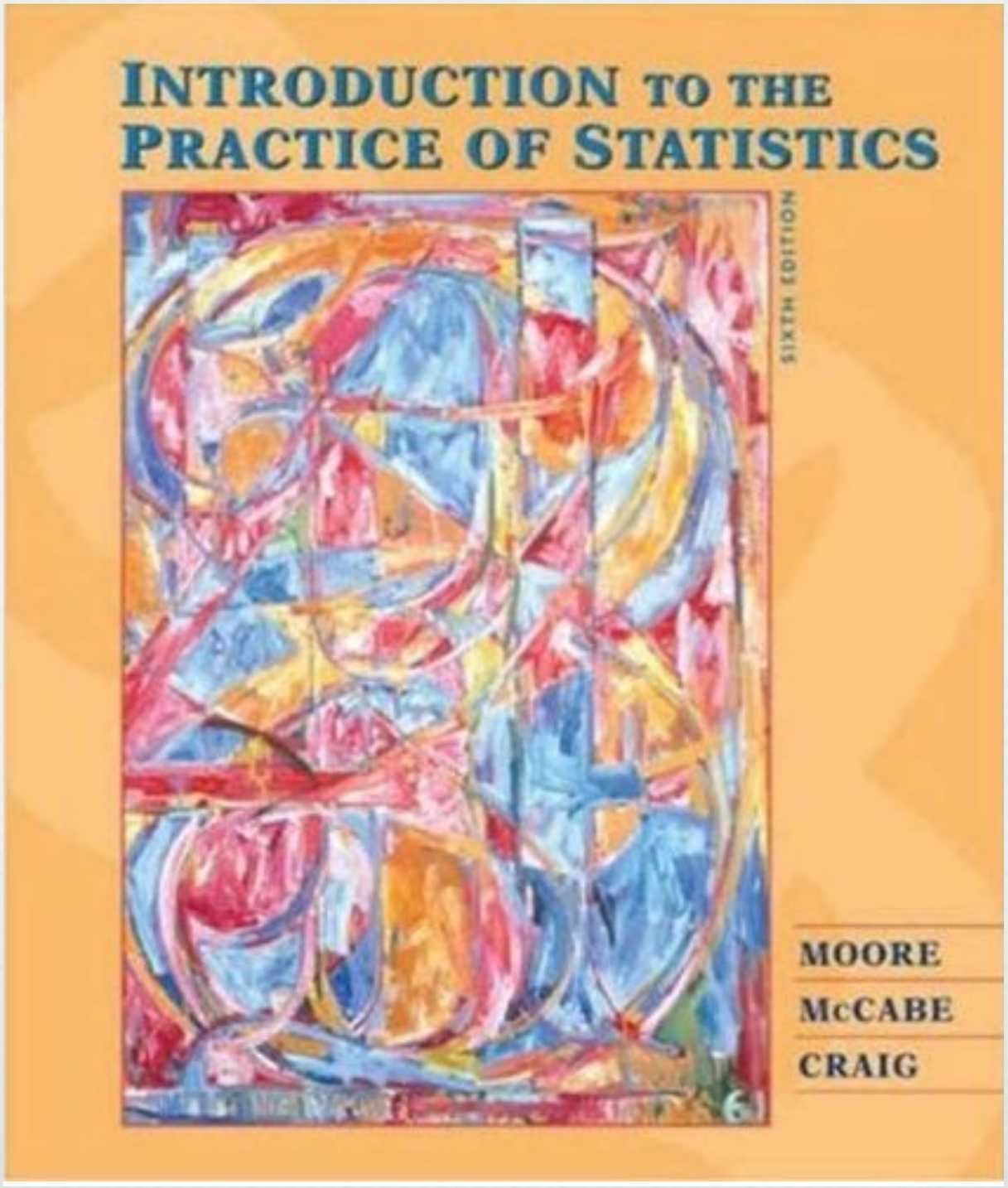


ALL OF THE CONTROLS WERE RUN BEFORE *ALL* OF THE CANCERS

THE MORAL OF THE STORY

- ☐ A BETTER MACHINE (QQTOF) WILL NOT SAVE YOU IF THE STUDY DESIGN IS POOR!
- ☐ THE ANSWER? RANDOMISE THE SAMPLE RUN ORDER!
- ☐ THERE IS NO WAY A WOMAN SHOULD BE TOLD SHE NEEDS SURGERY BASED ON THIS TEST!
- ☐ IN JUNE 2004, THE FDA RULED THAT OVACHECK COULD NOT BE MADE AVAILABLE UNDER THE "HOME BREW" EXEMPTION, AS THE SOFTWARE PROGRAM WAS A 'DEVICE' THAT NEEDED TO BE MORE TIGHTLY REGULATED.
- ☐ ... THESE RULES ARE BEING DEBATED EVEN NOW.
- ☐ AND WE ARE STILL WAITING FOR A VALID TEST FOR CANCER BASED ON BIOMARKERS!

DISASTER COULD HAVE BEEN AVOIDED ...



ACKNOWLEDGEMENTS

Keith Baggerly
M.D. Anderson Cancer Center
Texas

Terry Speed
Walter and Eliza Hall Institute
Melbourne