*Evolution of Statistical Science:*
*Translating Data to Innovative Health Care*

September 12–14, 2018

Washington Marriott Wardman Park

# Using Central Statistical Monitoring of Clinical Trials for Real-Time Detection of Data Anomalies

Marc Buyse, ScD
*marc.buyse@cluepoints.com*

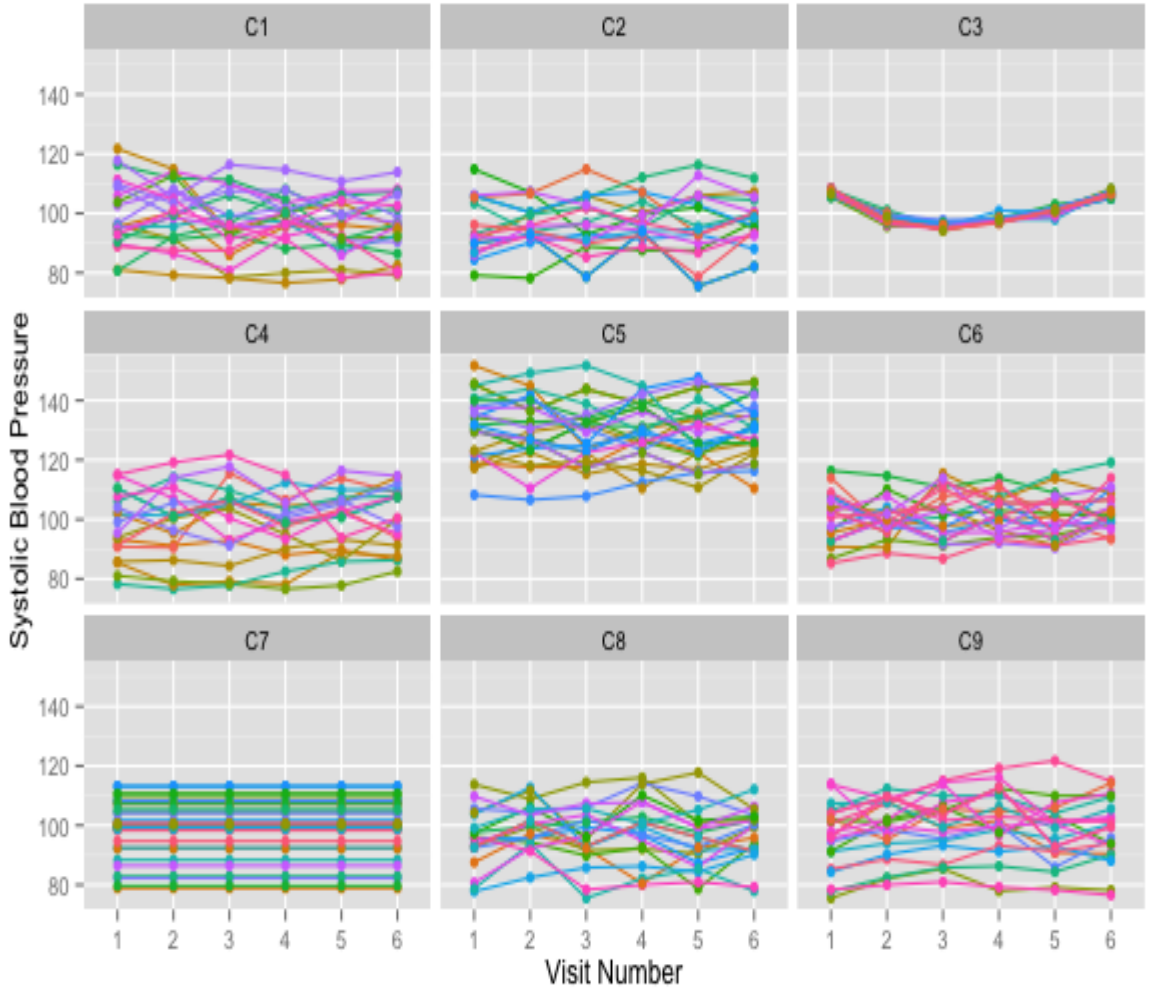
CluePoints

# Central Statistical Monitoring

Fabricated Data

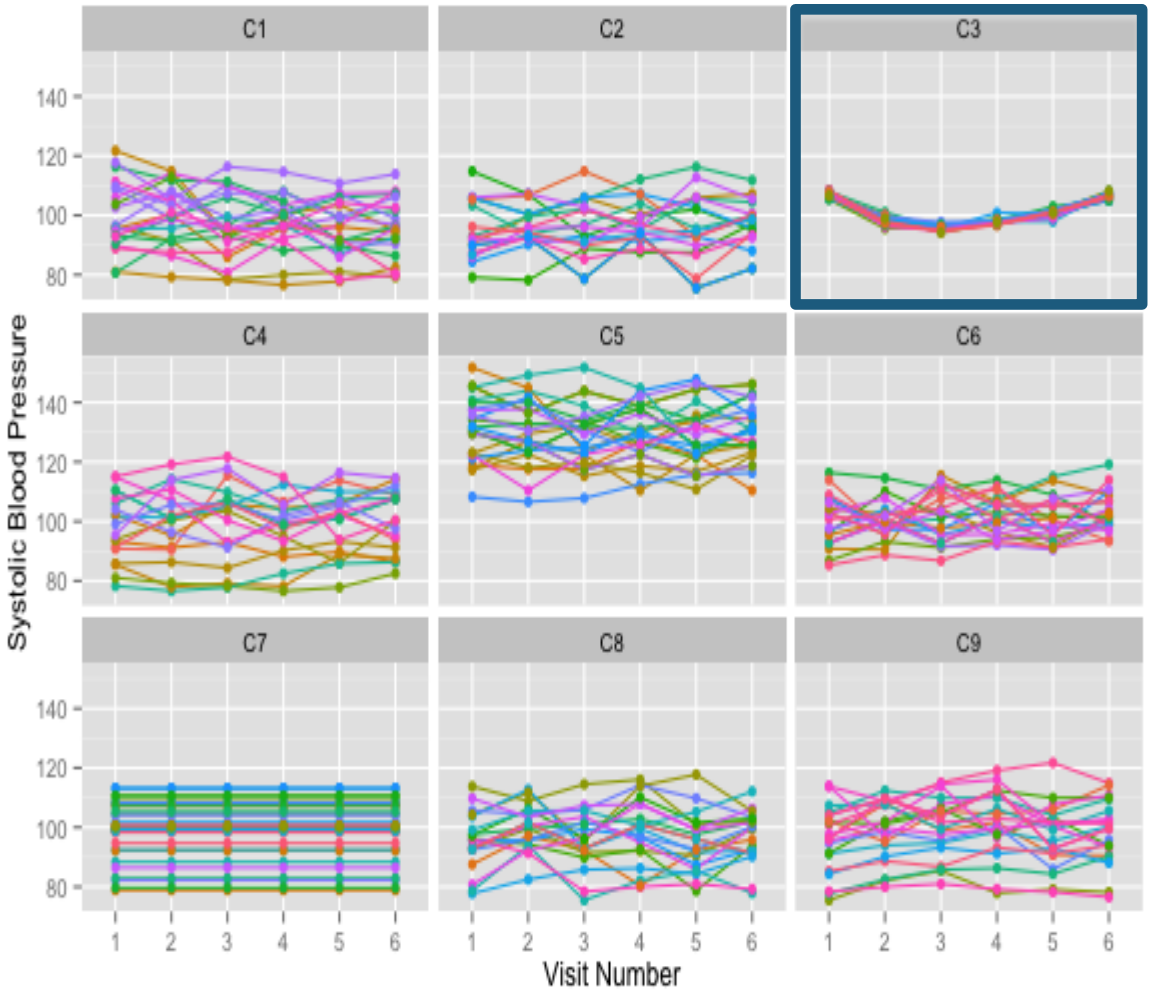
Actual Clinical Trial Data

# Data issues indicative of potential fraud



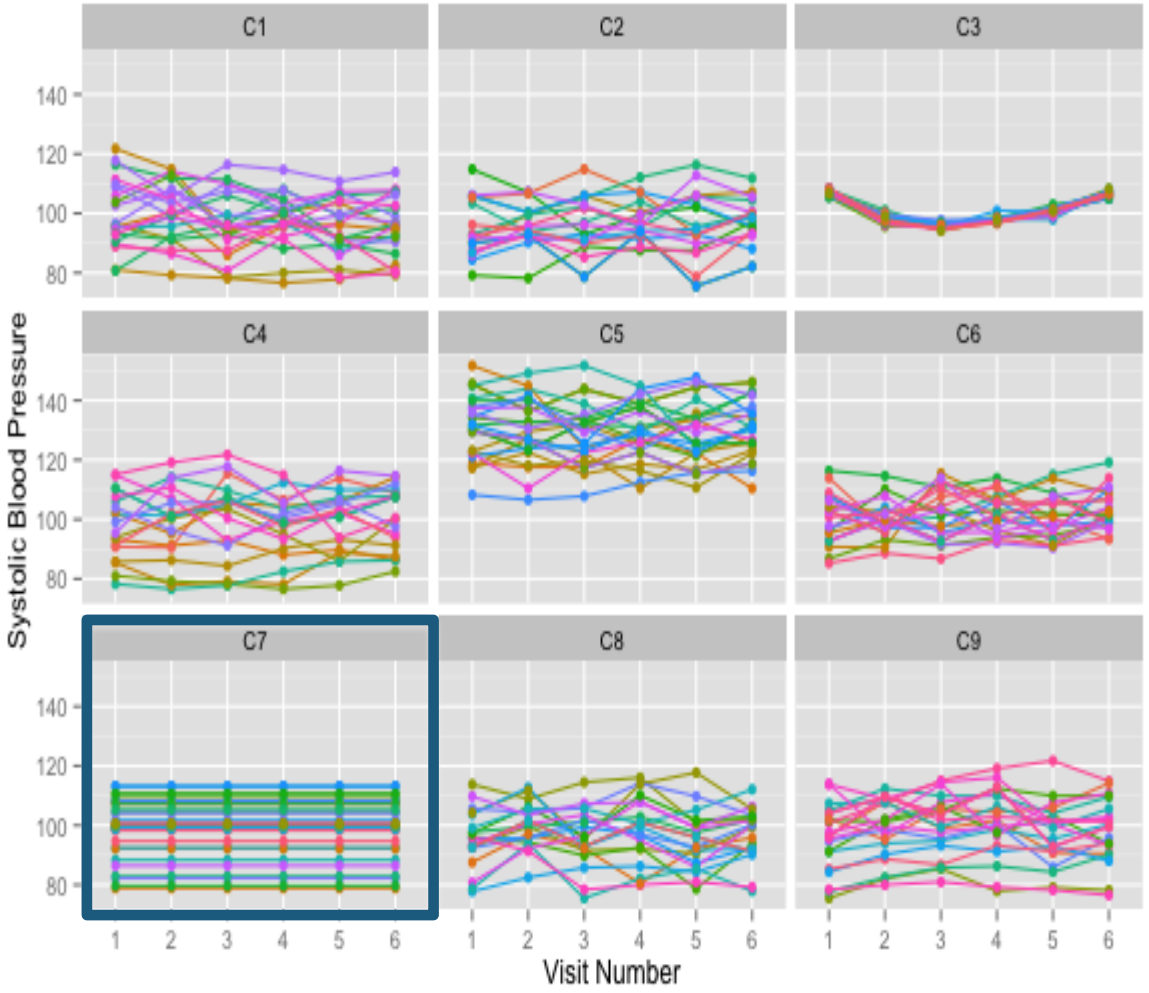
# Data issues indicative of potential fraud

## Lack of variability



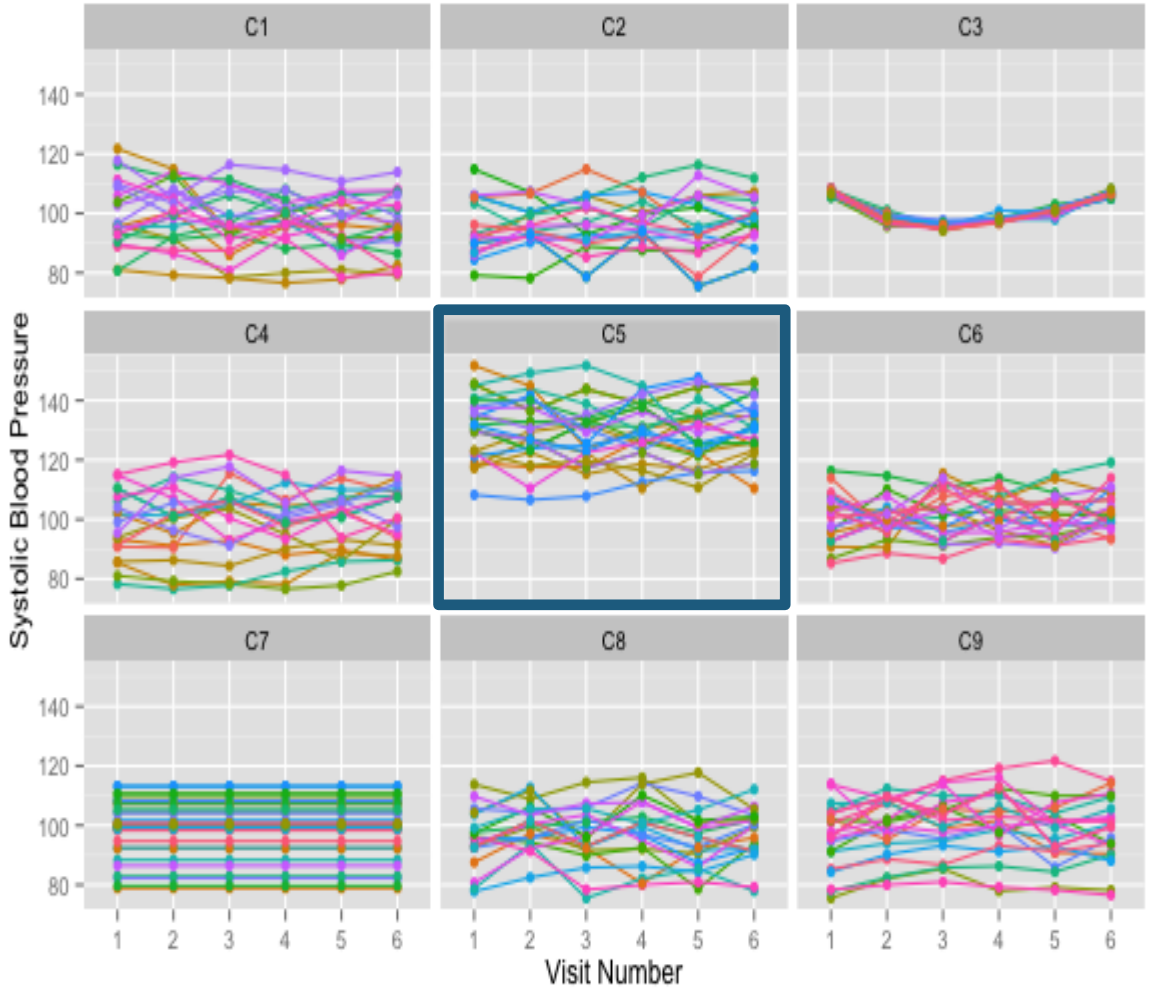
# Data issues indicative of potential fraud

## Data propagation



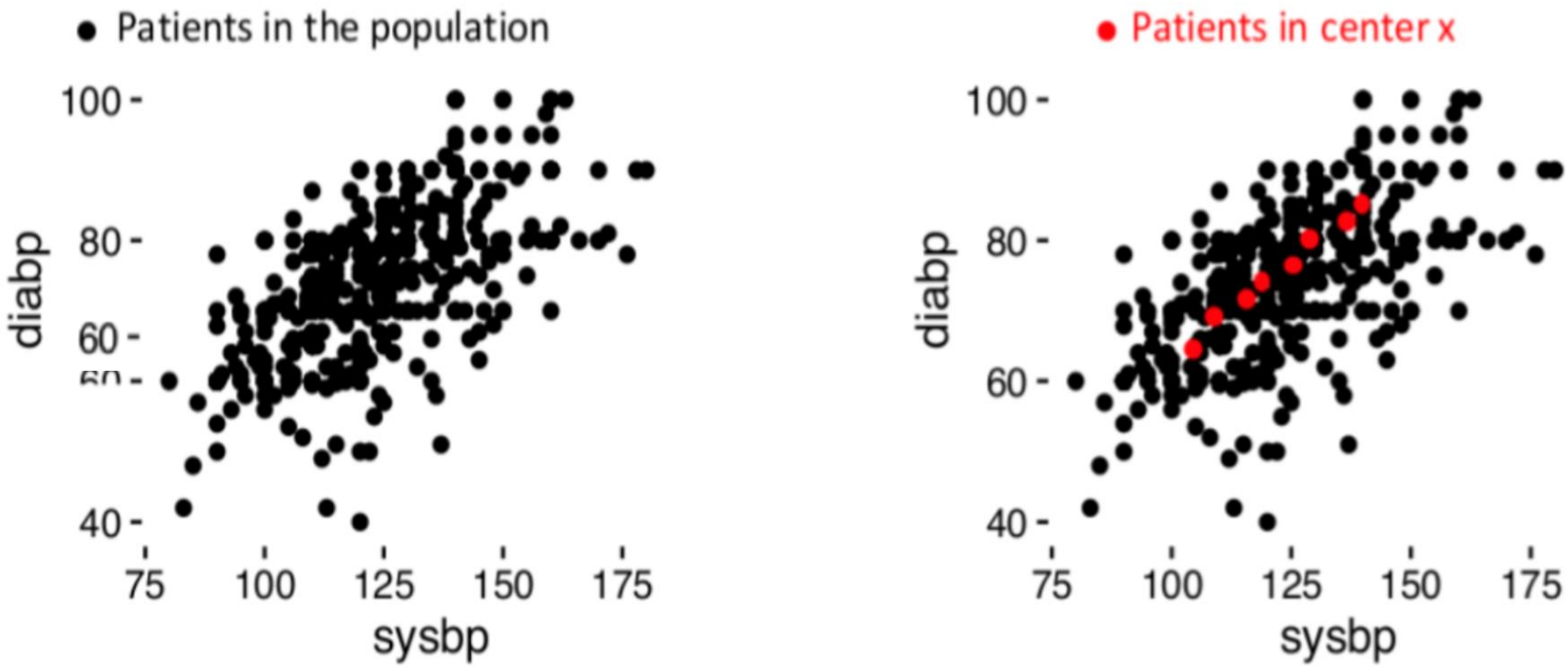
# Data issues indicative of potential fraud

## Shift in mean



# Data issues indicative of potential fraud

## Atypical correlation



# These ideas aren't new

STATISTICS IN MEDICINE
*Statist. Med.* **18**, 3435–3451 (1999)

## THE ROLE OF BIOSTATISTICS IN THE PREVENTION, DETECTION AND TREATMENT OF FRAUD IN CLINICAL TRIALS†

MARC BUYSE[1]*, STEPHEN L. GEORGE[2], STEPHEN EVANS[3], NANCY L. GELLER[4], JONAS RANSTAM[5], BRUNO SCHERRER[6], EMMANUEL LESAFFRE[7], GORDON MURRAY[8], LUTZ EDLER[9], JANE HUTTON[10], THEODORE COLTON[11], PETER LACHENBRUCH[12] AND BABU L. VERMA[13]

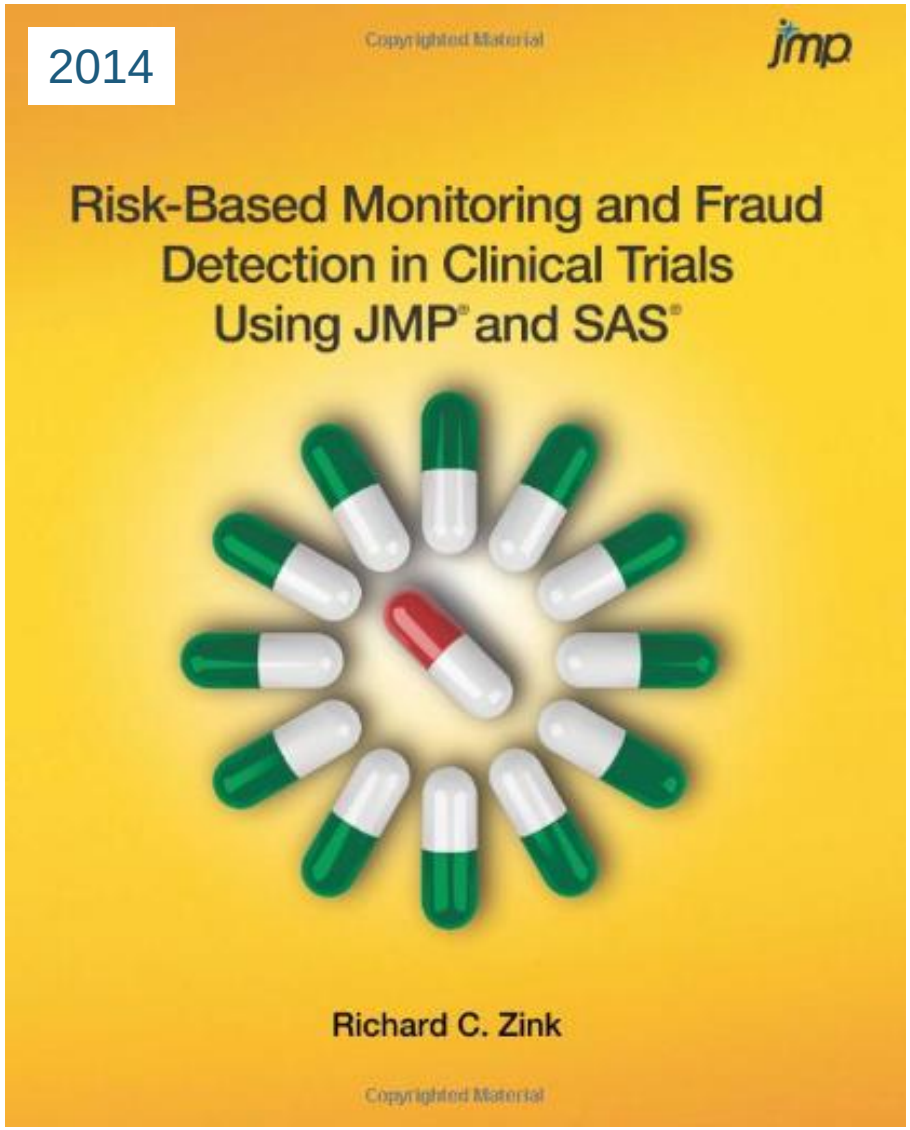
for the
ISCB SUBCOMMITTEE ON FRAUD

# These ideas aren't new

Table III. Some patterns that may reveal fraud in clinical trial data

| | |
|---|---|
| One variable at a time | Digit preference<br>Round number preference<br>Too few or too many outliers<br>Too little or too much variance<br>Strange peaks<br>Data too skewed |
| Several variables at a time | Multivariate inliers<br>Multivariate outliers<br>Leverage<br>Too weak or too strong correlation |
| Repeated measurements | Interpolation<br>Duplicates<br>Invented patterns |
| Calendar time | Breach of randomization<br>Days of week (Sundays or holidays)<br>Implausible accrual<br>Time trends |

# These ideas aren't new



# So, what _*is*_ new in CSM?

Central Statistical Monitoring (CSM) should be

- Unsupervised – no prior assumptions needed
- Agnostic – generic statistical tests not requiring any context knowledge
- Exhaustive – all data, regardess of importance, submitted to wide range of tests
- Robust – statistical tests require minimal distributional assumptions
- Sensitive and specific – use of mixed effects models to allow for natural variability

# CSM generates large matrix of *P*-values...

- *20 to 100 clinical sites*
- *250 to 1000 variables to test*
- *5 to 10 statistical tests per variable*

→ *$10^4$ to $10^6$ P-values*

| | A | B | DZ | EA | EB | EC | ED | EE | EF | EG | EH | EI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | test | mean | sdGlobal | mean | sd | sdGlobal | propagate | mean | sdGlobal | mean | sdGloba |
| 2 | | dataset | labs | labs | labs | labs | labs | labs | labs | labs | labs | labs |
| 3 | | variable | lab1 | lab1 | lab1r | lab1r | lab1r | lab1r | lab2 | lab2 | lab3 | lab3 |
| 4 | Center | Score | | | | | | | | | | |
| 5 | 1 | 0.0005 | -0.63 | 0.41 | -0.56 | | -0.33 | | -0.91 | -0.3 | -0.77 | 0.3 |
| 6 | 2 | 0.022 | 0.035 | -0.97 | 0.42 | -0.82 | 0.95 | 0.39 | -0.31 | 0.81 | 0.96 | 0.8 |
| 7 | 3 | 0.074 | -0.35 | -0.34 | -0.22 | -1 | -0.029 | 1 | -0.41 | 0.61 | -0.14 | 0.07 |
| 8 | 4 | 0.15 | -0.51 | -0.12 | -0.39 | 1 | -0.00068 | 1 | 0.7 | -0.041 | -0.75 | -0.7 |
| 9 | 5 | 0.27 | -0.47 | -0.19 | -0.78 | -0.81 | -0.29 | 0.62 | 0.88 | -0.61 | -0.46 | 0.3 |
| 10 | 6 | 0.27 | -0.99 | -3.7E-05 | -0.87 | -0.15 | -0.3 | 0.15 | 0.22 | 0.9 | 0.31 | -0.7 |
| 11 | 7 | 0.33 | -0.87 | -0.82 | 0.68 | | 0.71 | | 0.32 | -0.19 | 0.15 | -0.8 |
| 12 | 8 | 0.48 | 0.6 | 0.86 | 0.59 | | -0.095 | | 0.68 | -0.039 | 0.45 | 0. |
| 13 | 9 | 0.52 | -0.95 | 0.9 | -0.86 | 0.74 | 0.89 | 0.46 | -0.88 | -0.041 | 0.94 | -0.8 |
| 14 | 10 | 0.71 | 0.98 | -0.16 | 0.89 | -0.11 | -0.32 | 0.11 | 0.62 | -0.7 | 0.52 | 0.1 |
| 15 | 11 | 0.78 | -0.94 | -0.013 | -0.21 | | -0.058 | | 0.18 | -0.41 | 0.28 | -0.2 |

# ... and a data consistency score* for each site

Table 1: Interpretation of scores ($S = -log_{10}(P)$)

| Score ($S$) | Interpretation | Expected frequency under the null |
|---|---|---|
| $S < 1.3$ | No statistical signal | 0.95 |
| $1.3 \leq S < 2.0$ | Minor statistical signal possibly due to chance | 0.04 |
| $2.0 \leq S < 4.0$ | Moderate statistical signal requiring attention | 0.0099 |
| $4.0 \leq S < 6.0$ | Strong statistical signal requiring further scrutiny | 0.000099 |
| $S \geq 6.0$ | Extreme statistical signal requiring detailed investigation | 0.000001 |

Charles Francis Richter

** Data inconsistency score for N tests = $[\Sigma_{i=1,N} -log(P_i)]/N$*

# Fraud is only one type of issue

| Type | Typical examples | Intent |
|---|---|---|
| Fraud | Falsified data<br>*(e.g. fabrication of ePRO data)* | Intention to cheat |
| Tampering | Fabricated data<br>*(e.g. propagation of blood pressure)* | Deliberate |
| Sloppiness | Incorrect reporting<br>*(e.g. under-reporting of AEs)* | Limited awareness |
| Errors | Technical problems<br>*(e.g. miscalibrated thermometers)* | Unintentional |

Central Statistical Monitoring

Fabricated Data

Actual Clinical Trial Data

# Simulation setting (1)

Data fabricated by 34 applied scientists (height & weight of 40 hypothetical students), based on known distribution

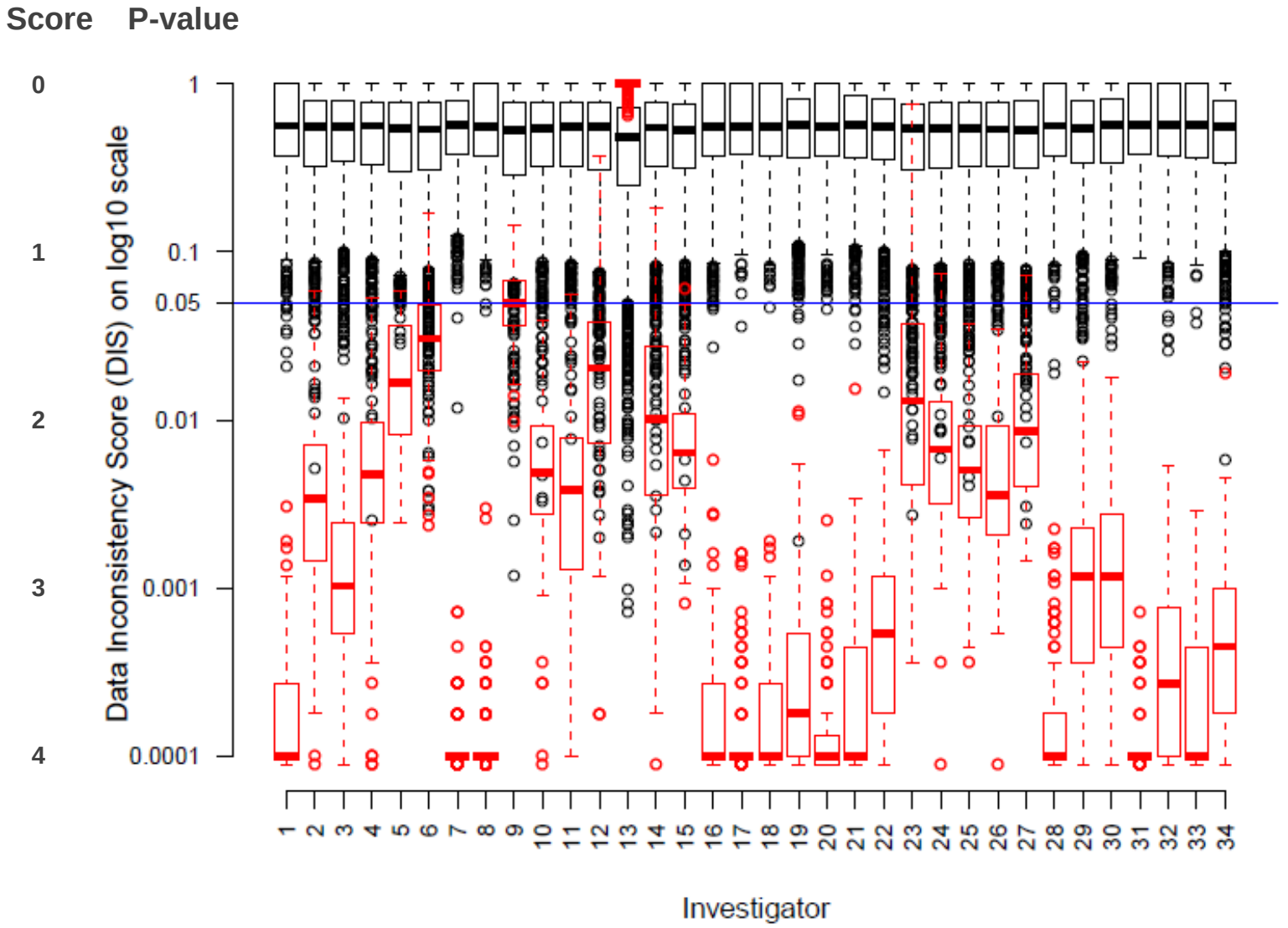
| Variable | Minimum | Maximum | Mean | Standard deviation | Correlation |
|---|---|---|---|---|---|
| **Weight (kg)** | 39 | 84 | 54.5 | 9.2 | r = 0.43 (P < 0.001) |
| **Height (cm)** | 145 | 175 | 159.5 | 7.2 | |

# Simulation setting (2)

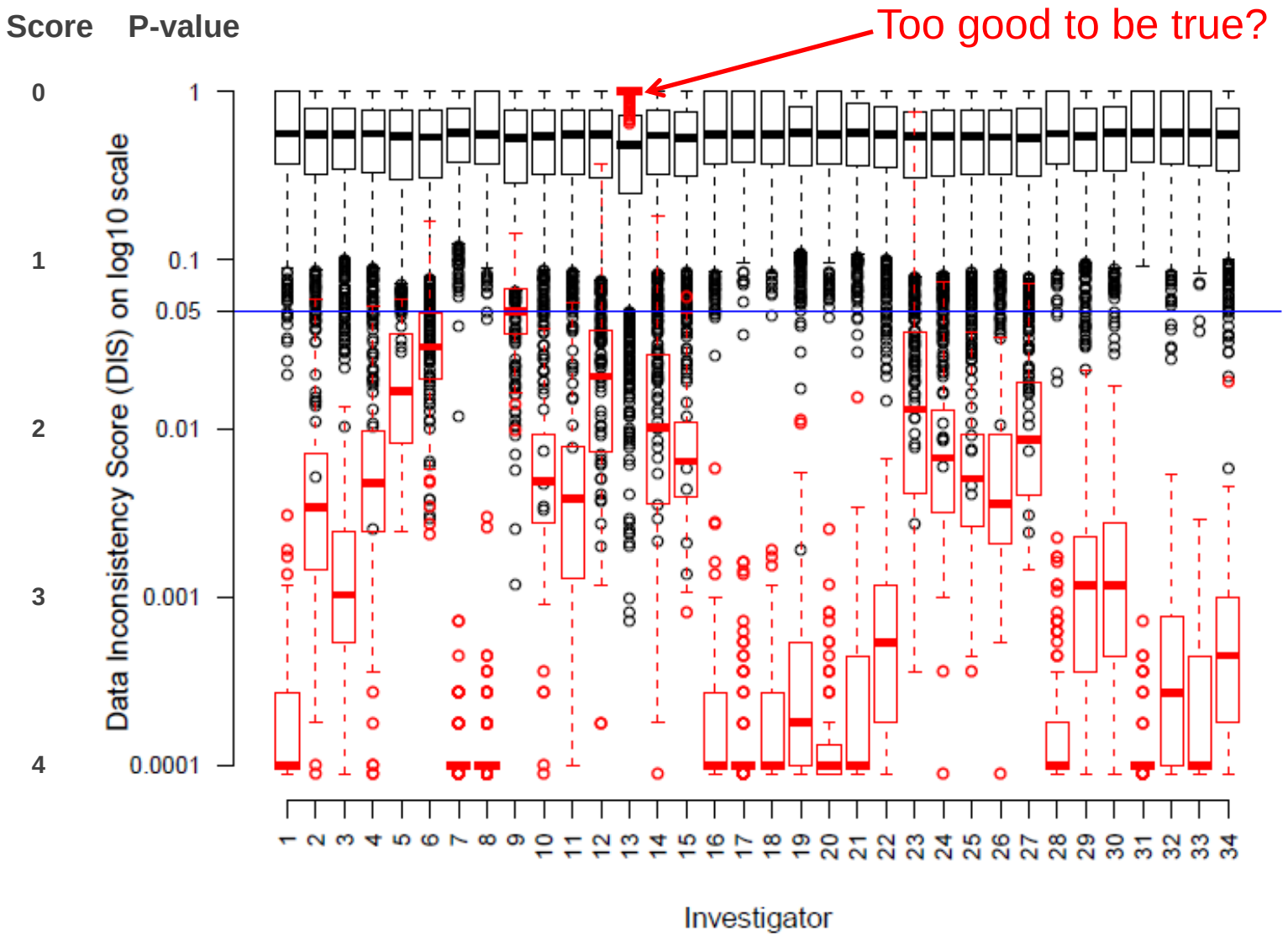
Simulations:

- Generate 100 trials with 20 centers of 40 patients each, 19 « genuine » centers (height and weight sampled from known distribution) and 1 « fraudulent » center (heights and weights fabricated)
- For each trial, calculate the data inconsistency score of all centers
- Detect outlying centers (large data inconsistency scores)

# Sensitivity of data inconsistency score



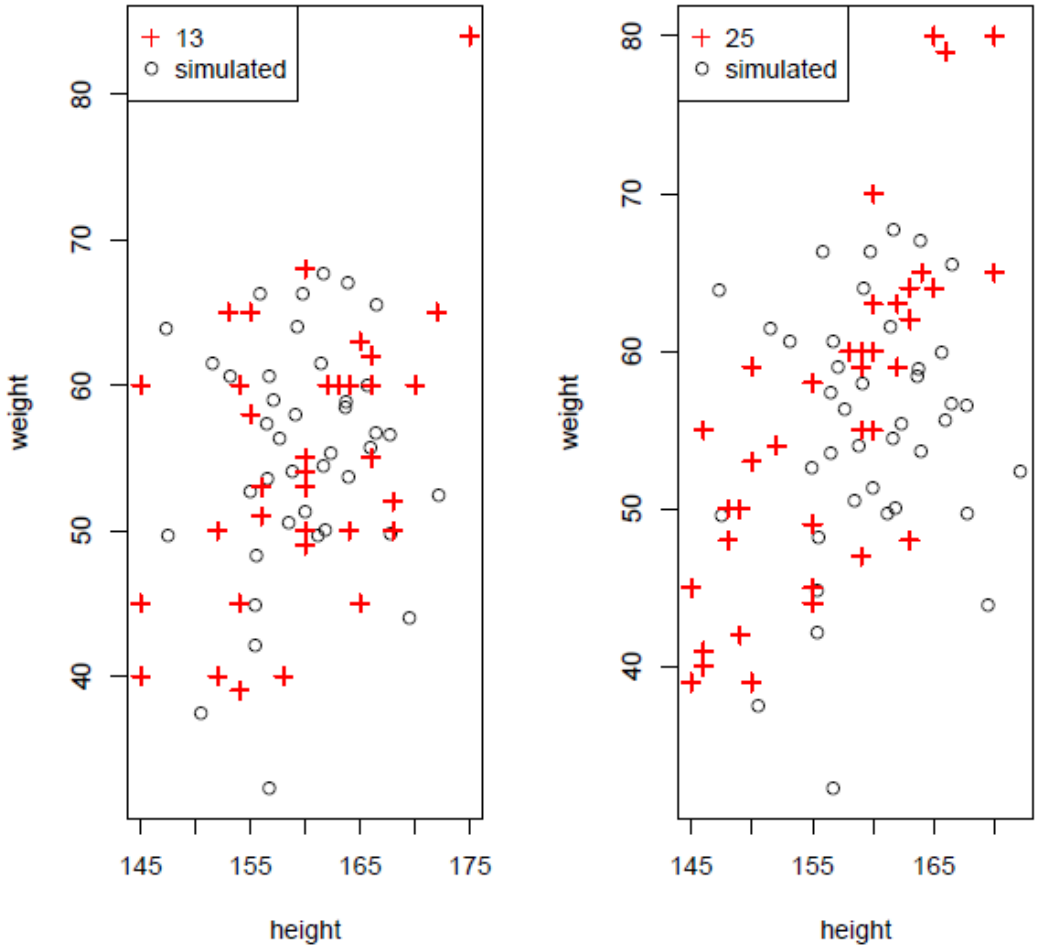
# Sensitivity of data inconsistency score



# Conclusions: data fabrication is detectable

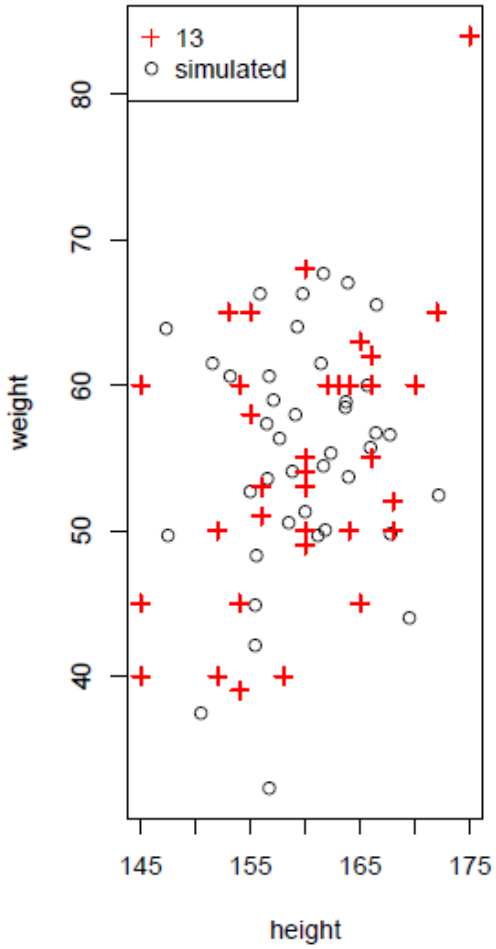
- People (including scientists) cannot invent plausible data
- Computer algorithms can (investigator #13)
- CSM has high sensitivity (and specificity) for the detection of most fabricated data

# Data visualization

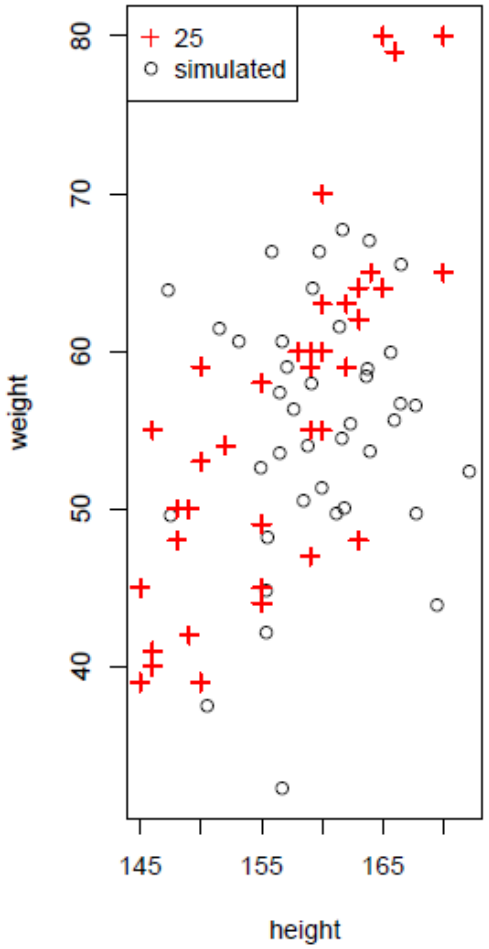


# Data visualization

Too good to be true



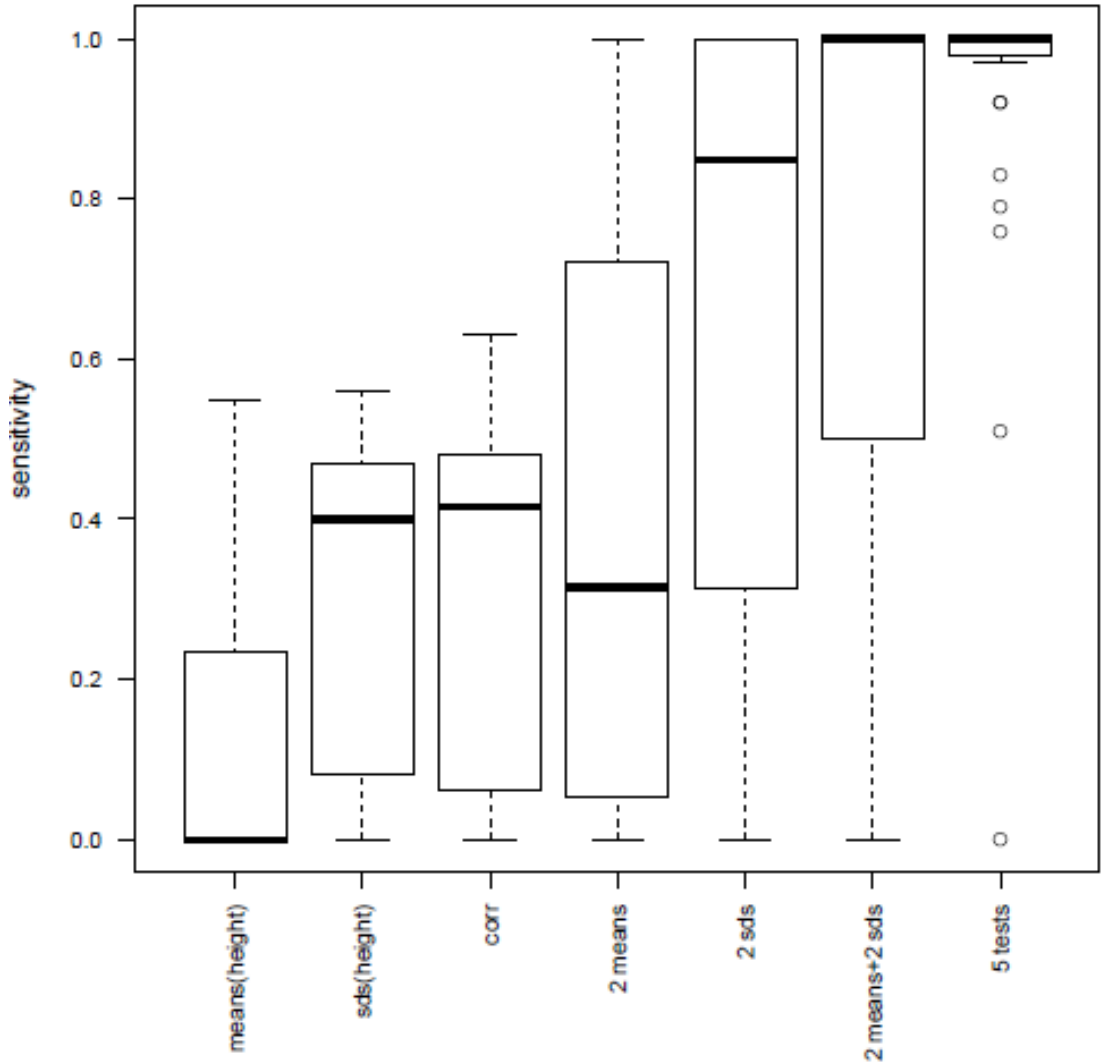

Detected as fabricated



# Conclusions: visual inspection is not sufficient

- People (including scientists) cannot invent plausible data
- Computer algorithms can
- CSM has high sensitivity (and specificity) for the detection of most fabricated data
- Data visualization tools are generally not sufficient

# Sensitivity with increasing number of tests



# Conclusions: one variable is not sufficient

- People (including scientists) cannot invent plausible data
- Computer algorithms can
- CSM has high sensitivity (and specificity) for the detection of most fabricated data
- Data visualization tools are generally unhelpful
- Sensitivity increases dramatically when the number of variables / tests increases

Central Statistical Monitoring
Fabricated Data

# Actual Clinical Trial Data

# Simulation setting (1)

Used data from actual clinical trial conducted in Japan

Data contamination:

**Table 1.** Parameters for data contamination

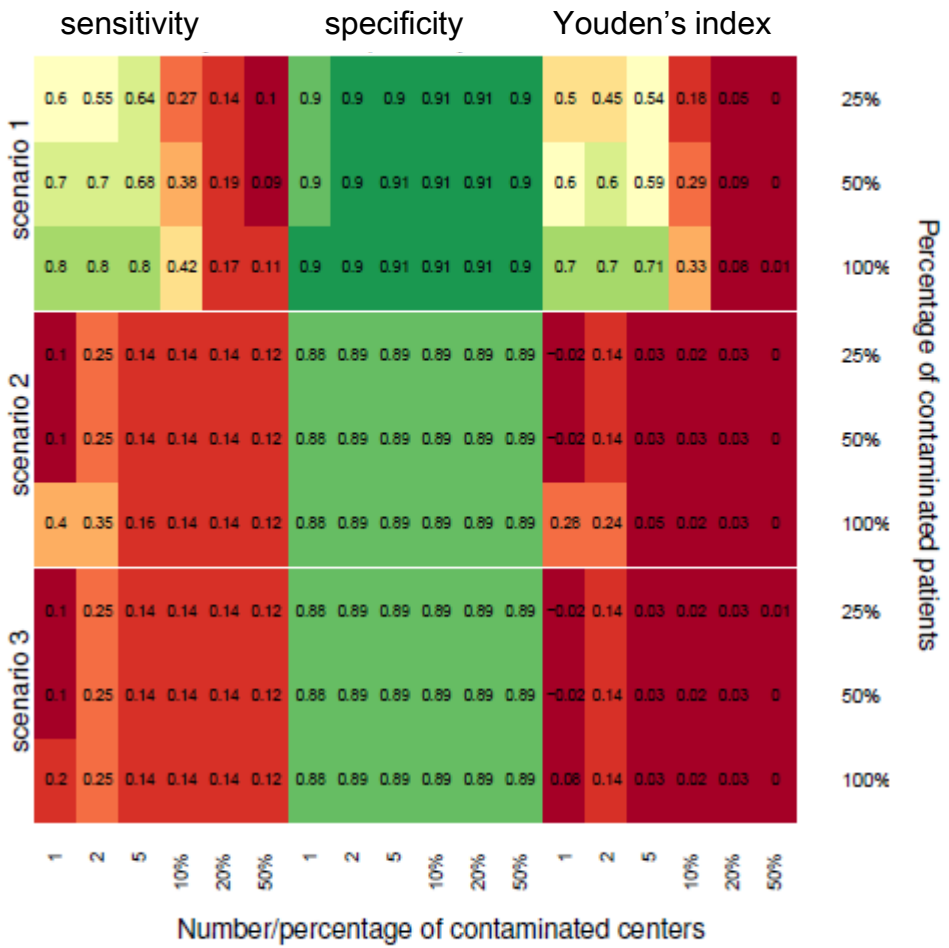
| Parameter | Values | Number of patterns |
|---|---|---|
| Number or proportion of contaminated centers | 1, 2, 5, 10%, 20%, 50% | 6 |
| Proportion of patients contaminated by center | 25%, 50%, 100% | 3 |
| Number and types of variables contaminated | See Scenarios in Table 2 | 9 |
| Total number of patterns | | 162 |

# Simulation setting (2)

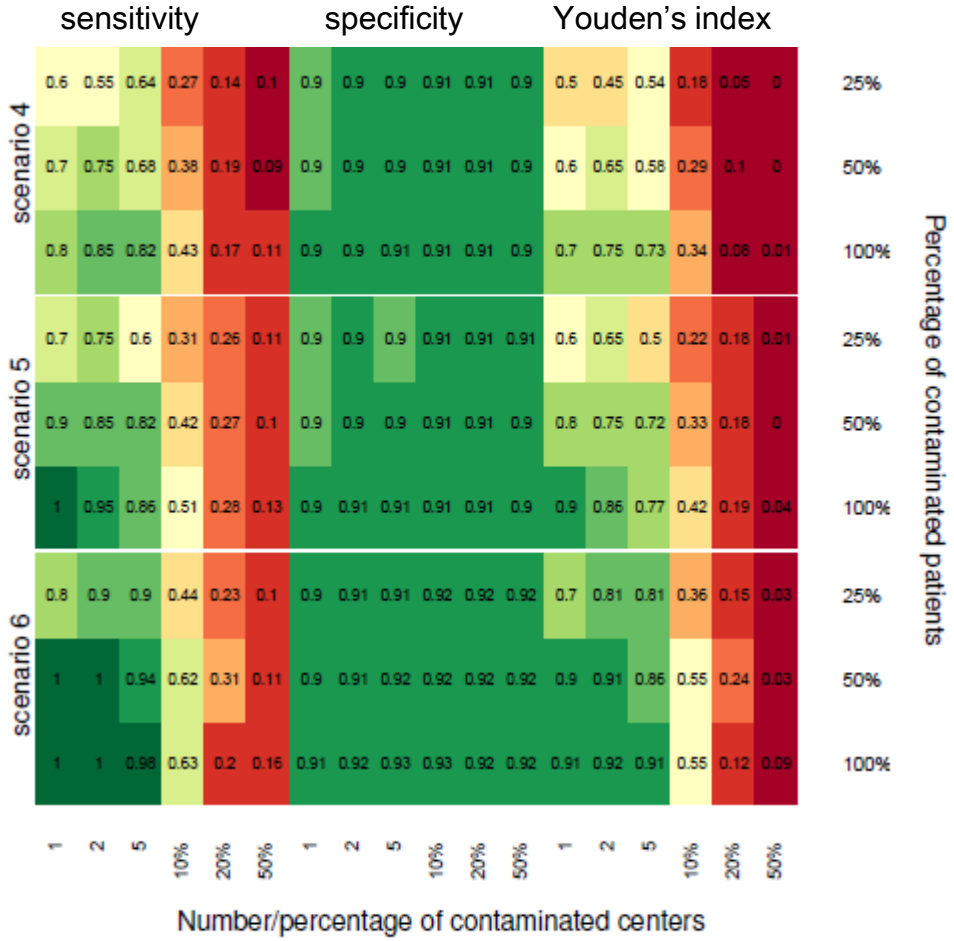
**Table 2.** Scenarios for data contamination

| | | Variable type | | |
|---|---|---|---|---|
| | | Continuous | Categorical | Date |
| One variable contaminated | Scenario 1 | 1 | 0 | 0 |
| | Scenario 2 | 0 | 1 | 0 |
| | Scenario 3 | 0 | 0 | 1 |
| Several variables contaminated | Scenario 4 | 1 | 1 | 1 |
| | Scenario 5 | 2 | 2 | 2 |
| | Scenario 6 | 3 | 3 | 3 |
| Proportion of variables contaminated | Scenario 7 | 5% | 5% | 5% |
| | Scenario 8 | 10% | 10% | 10% |
| | Scenario 9 | 20% | 20% | 20% |

| | | Variable type | | |
|---|---|---|---|---|
| | | Continuous | Categorical | Date |
| One variable contaminated | Scenario 1 | 1 | 0 | 0 |
| | Scenario 2 | 0 | 1 | 0 |
| | Scenario 3 | 0 | 0 | 1 |



| | | Variable type | | |
|---|---|---|---|---|
| | | Continuous | Categorical | Date |
| Several variables contaminated | Scenario 4 | 1 | 1 | 1 |
| | Scenario 5 | 2 | 2 | 2 |
| | Scenario 6 | 3 | 3 | 3 |



| | | Variable type | | |
|---|---|---|---|---|
| | | Continuous | Categorical | Date |
| Proportion of variables contaminated | Scenario 7 | 5% | 5% | 5% |
| | Scenario 8 | 10% | 10% | 10% |
| | Scenario 9 | 20% | 20% | 20% |

# Conclusions

- The data inconsistency score has high specificity
- The data inconsistency score has higher sensitivity when
  - a smaller proportion of centers have data issues
  - the number of variables affected increases
  - variables affected are continuous
  - variables are repeatedly measured over time

# Some references


- Buyse M, *et al*. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* 1999; 18: 3435-51.
- Venet D, *et al*. A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials* 2012; 9: 705-13.
- George S, Buyse M. Data fraud in clinical trials. *Clinical Investigation* 2015; 15:161-73.
- Buyse M, Evans SJW. Fraud in clinical trials. In: *Wiley StatsRef: Statistics Reference Online,* 2016, John Wiley & Sons. DOI 10.1002/9781118445112.stat04938.pub2
- Trotta L, *et al*. Unsupervised statistical monitoring for the detection of atypical data in multicenter clinical trials. 2018 (Submitted).


*Further questions? marc.buyse@cluepoints.com*