DATA ANALYSIS AND REPORTING

Biostatistical Challenges in R&D

Conflicting regulators, upbeat developers and big data: How to bring them together?

Gonnie van Osta

Author! et al. BV

Introduction

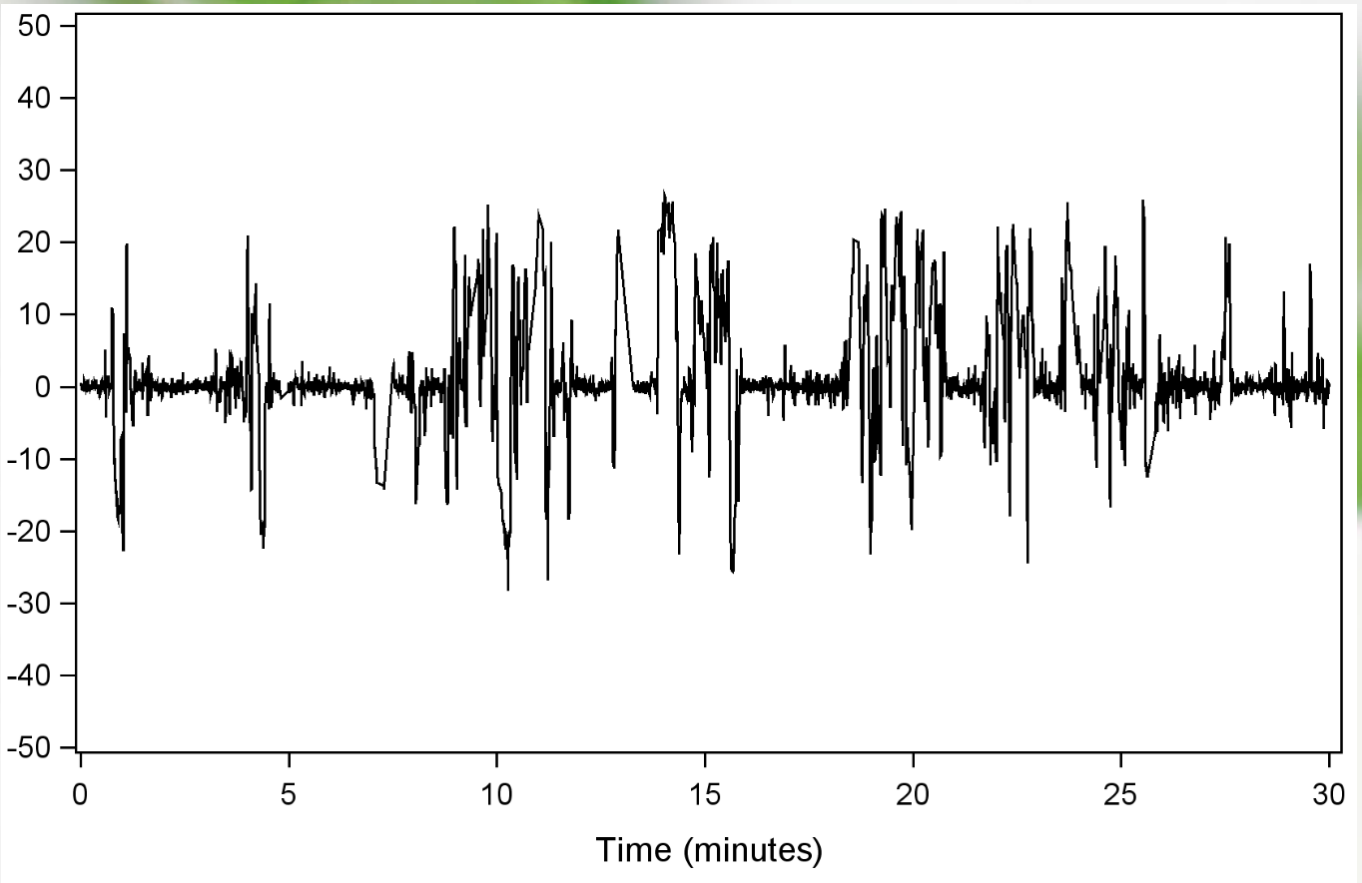
- Gonnie van Osta (Goes, 1962)
- First year, Human Movement Studies VU, 80s
- MSc in Mathematical Statistics UvA, 80s
- 3 years, statistical consultant DLO Wageningen
- 22 years in development (biometrics, qualitiy, clinical, regulatory, pharmaceutical) Organon etc, Oss
- Registered biostatistician, 2000
- Scientific meeting organisator PSDM/EFSPI, 2002-2006
- Lean six sigma black belt, 2012
- Currently: statistical consultant at AUTHOR!

Example: Diagnostic Medical Device

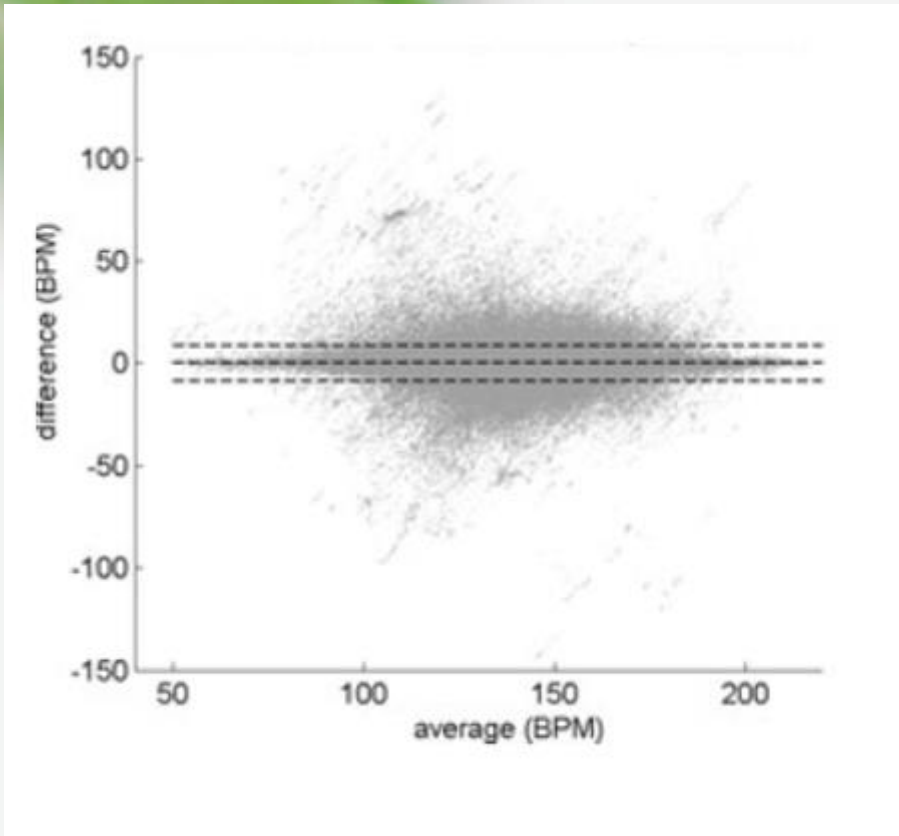
- New device to measure heartbeat, less invasive
- Aim: to replace the existing device with the new device
- Request: Study design/power calculations to show that the new device is as good as the golden standard
- What is measured?
  - 2 Devices in parallel (paired)
  - Heartbeat (periferal), in various stages of physical effort
  - Periods: several hours
  - 4 observations per second

→ Lots and lots of data

The data, one patient,±7000 points



Indication literature

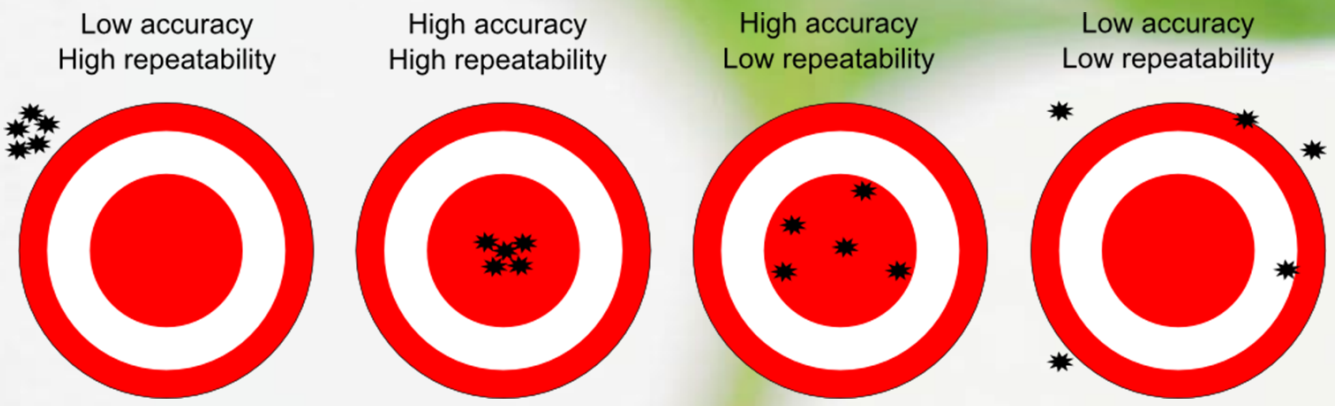


Challenge

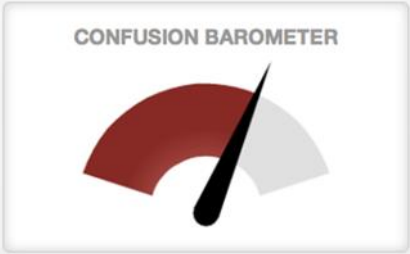
- Input: sponsor, indication literature, hospital EC, regulators
- Sponsor/Literature:
  - Literature, 3 arm study showing superiority of one new devices over another existing device.
  - Reliability=Percentage Positive Agreement=Percentage Time Heart Beat of 2 systems is within 10 beats
  - Accuracy: root MSE of differences (or against the regression of Bland-Altman plot?)
  - 3-arm study not feasible: non-inferiority 2-arm

Aim for a reliability and accurate method

- Reliability=Percentage Positive Agreement=Percentage Time Heart Beat of 2 systems is within 10 beats
- Accuracy: SD estimation of paired differences
- Literature: Greenwood 1950: Sample Size Required For Estimating The Standard Deviation as a Percent of Its True Value, used for military (seemed appropriate), N=80



Challenge



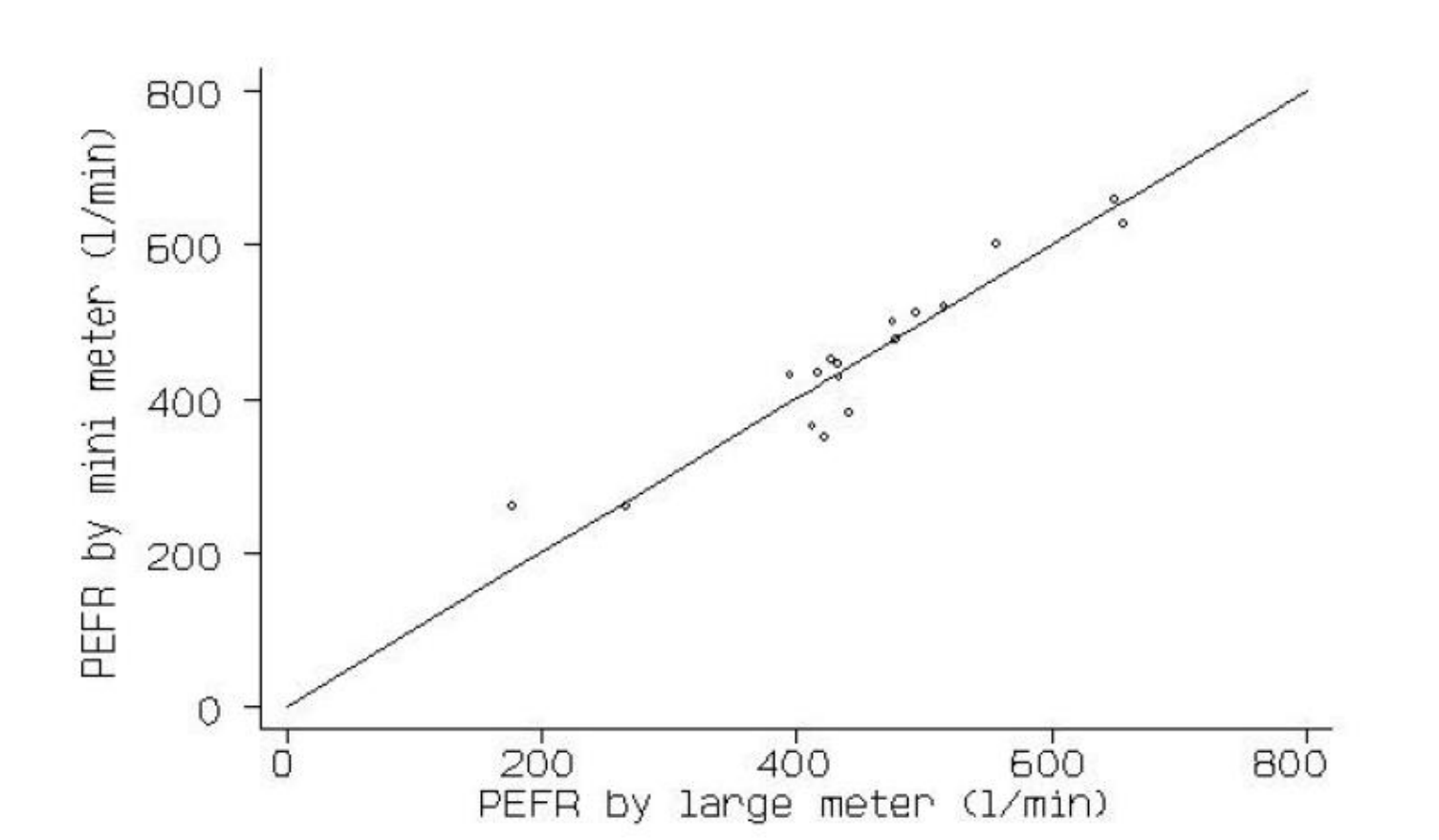
- Regulators, show reliability and accuracy against golden standard:
  1. Reliability and Accuracy: N=80 seems low, use Bland/Altman 1983 to determine sample size for limits of agreement and bias estimation
  2. Reliability: Proposed definition of reliability is loss of information and repeated measures, use Deming regression (β_0=0, β_1 =1).
  3. Accuracy: there are correlated repeated measures, use bootstrapping methods when constructing CIs for bias, Bland-Altman (2007) analysis including plots.

Limits of agreement is the new definition of reliability.

What is this new definition?

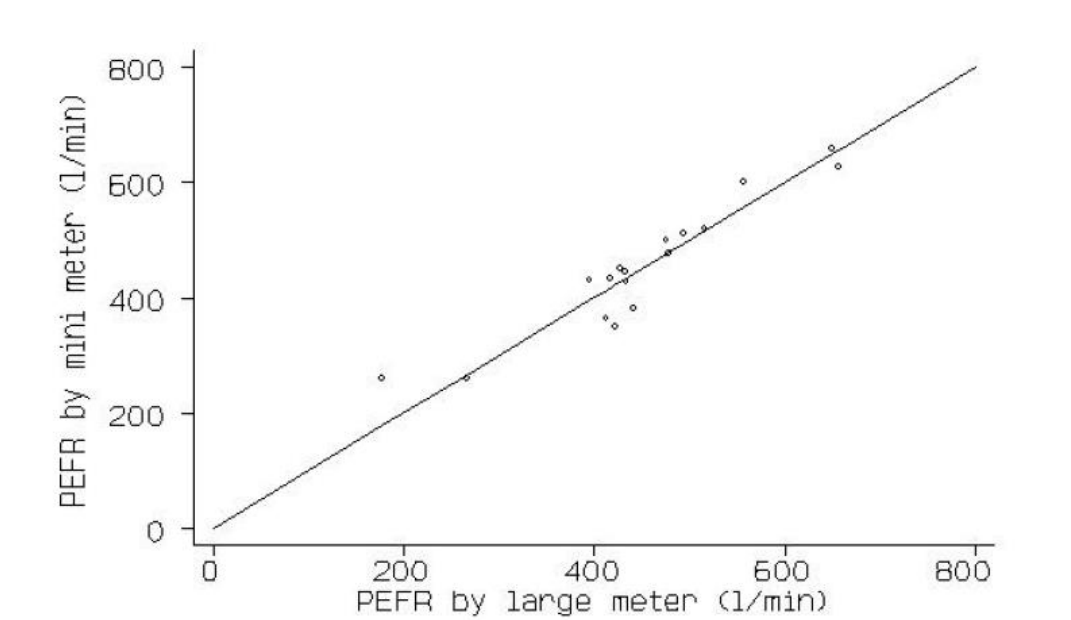
Bland & Altman, Agreement between methods of measurement with multiple observations per individual. Journal of Biopharmaceutical Statistics, 17: 571–582, 2007

Bland-Altman (1983) side-step



Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307–17

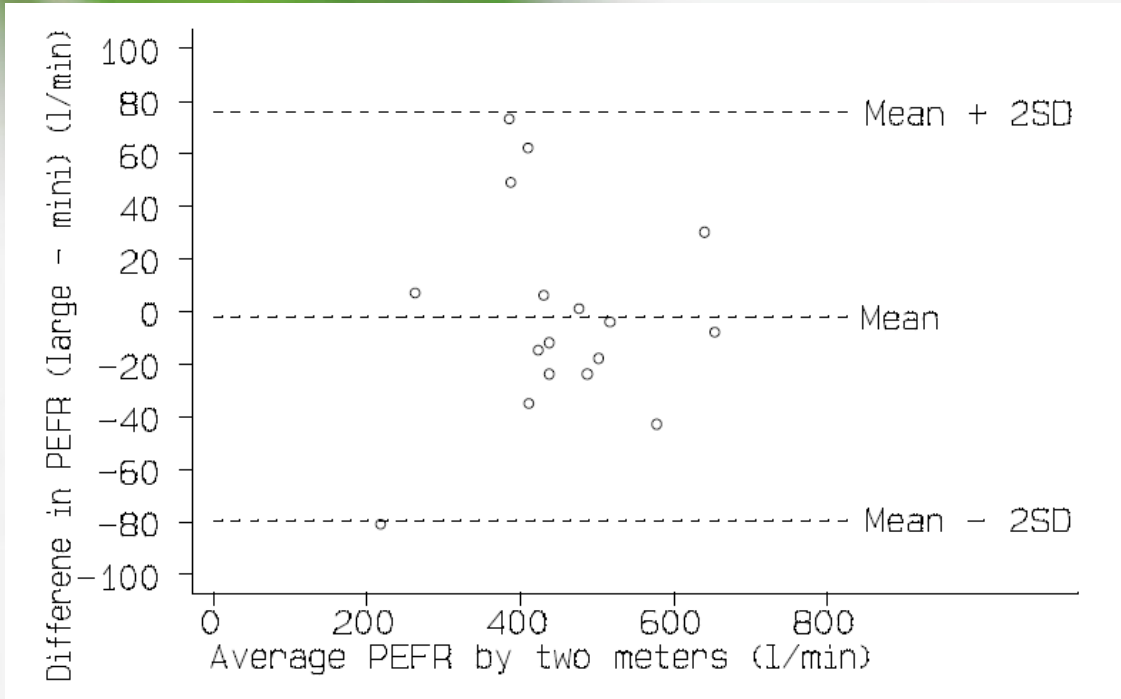
Bland-Altman (1983) side-step



- Data will cluster around a regression line
- The greater the range of measurements the greater the agreement will appear to be.

→ regression is not the way

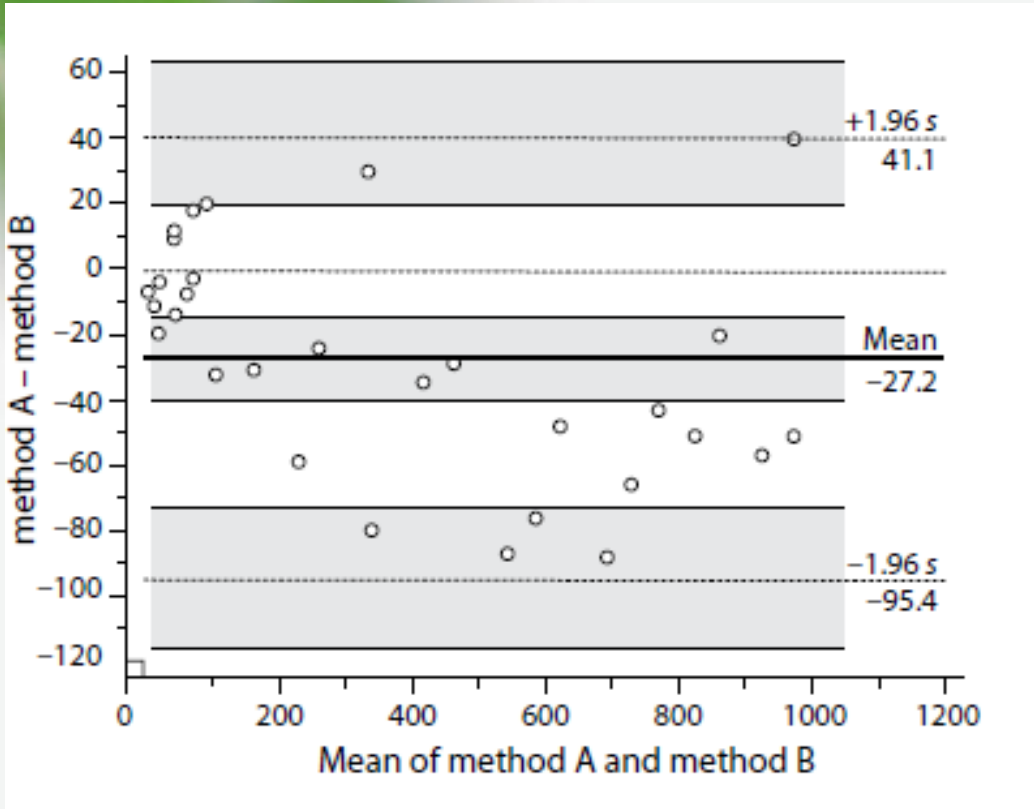
Bland-Altman (1983) side-step



Bland-Altman plot:

- Difference against average
- Error and bias are much easier to assess
- Bias -2.1, mean +/- 2*SD ranges from -80 to +76, this lack of agreement not clear from regression figure

Bland-Altman: Limits of agreement



Bland-Altman plot (continued):

- Estimation of precision of SD/limits of agreement depends on sample size, SE of limits is sqrt(3S**2/n)

Giavarina (2015), Lessons in Biostatistics: Understanding Bland Altman analysis

Bland-Altman side-step

Conclusions:

- Correlation does not measure agreement
- Least square regression does not measure comparibility
- This is not callibration. Since callibration is the situation where the true value is known

Summary/Assumptions:

- Paired (single readings)
- Uncorrelated
- Repeatability/plots: Investigate the between method differences and relation with the size of the measurements

Example: Diagnostic Medical Device

So far, straightforward, use Bland-Altman.

- But which one? 1983 or 2007?

In the mean time:

- Trouble managing the large amounts of data
- Lots of (test) data
  - Not keen on bootstrapping
  - Plotting to check B&A assumptions is a challenge
  - Deming regression (β_0=0, β_1 =1) or Bland-Altman (dif vs average regression)?
  - Accounting for correlated repeated data

Bland-Altman side-step

Our example

- Paired observations ✓
- Independent observations **X**
- No relation between difference(bias) and mean **?**

Example: Diagnostic Medical Device

Our test data:

- Independent: **X**
- Relation Bias and mean **?**

- Bland&Altman 1999/2007:
  - Number of obs per patient varies (2-5)
  - True value varies

  → One way analysis, estimate residual mean square (1 summary per patient).

But: observations within a patient are assumed independent

Example: Diagnostic Medical Device

Our test data:

- Independent: **X**
- Relation Bias and mean **?**

- Dependency

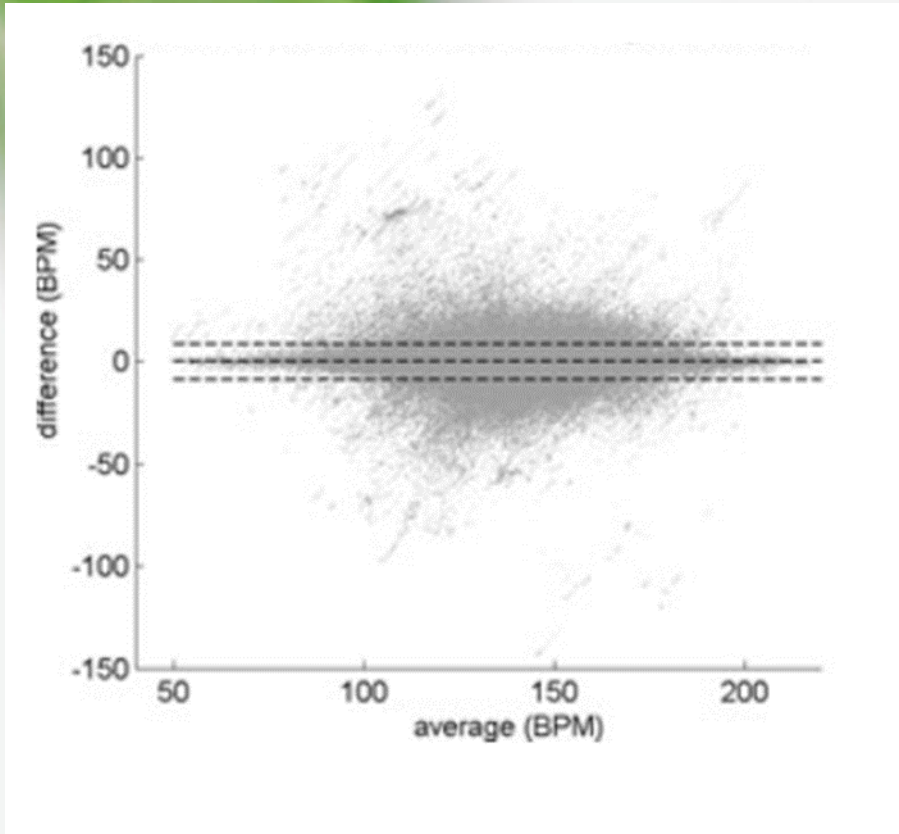
  → Estimate correlation or use only one data-point?

Hours*minutes*seconds*4 >100.000 paired observations per patient

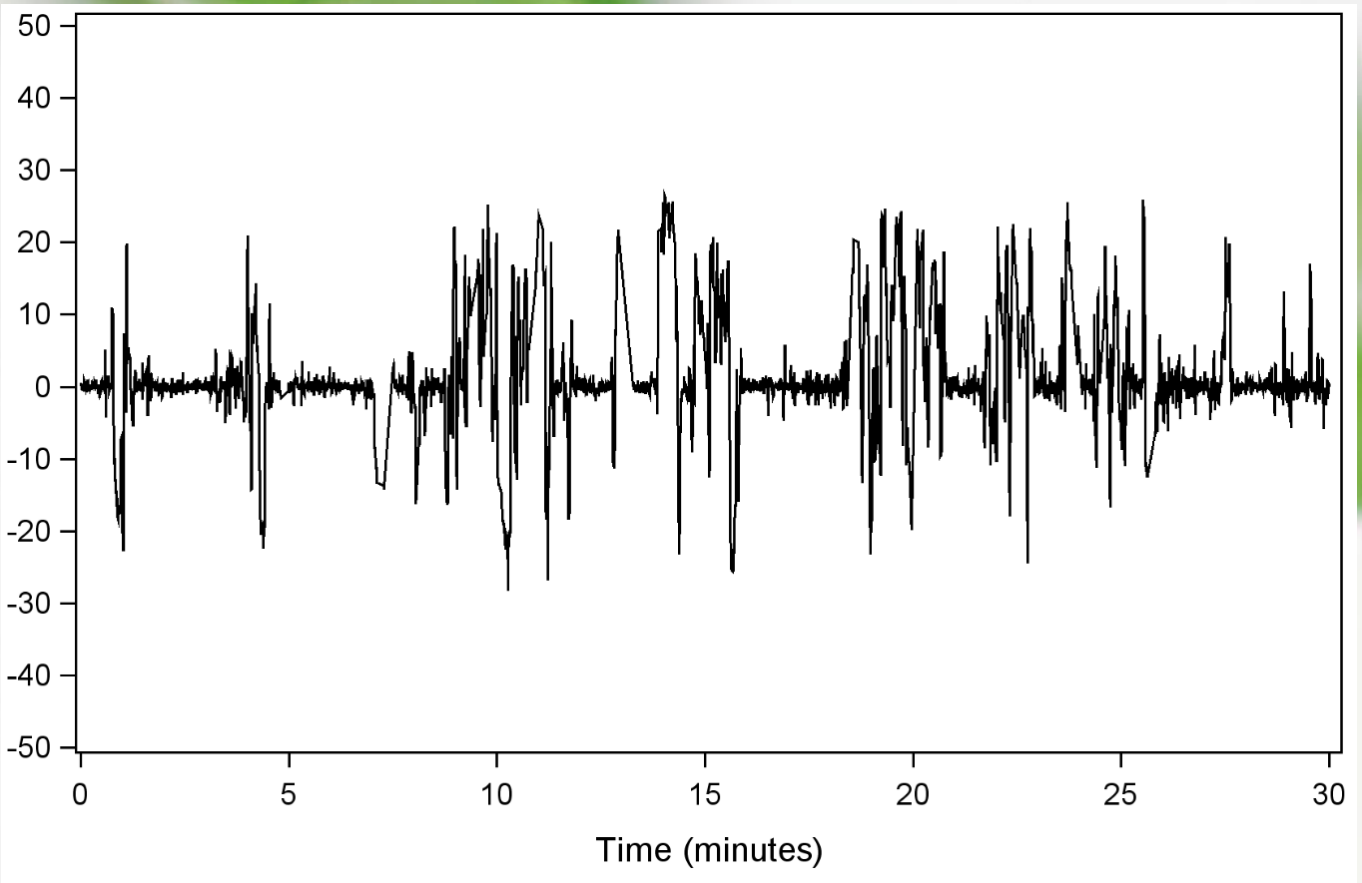
→ Hard to estimate/model correlation

→ Hard to explore graphically (B&A plot or Regression plot)

Literature: Bland-Altman plots



The data, one patient,±7000 points



Example: Diagnostic Medical Device

- Our final data (average ~55.000 paired points per patient)

How can it be that I am longing for fewer data-points?

So, what did we do?

For regulators that were not concerned with repeated measures:

- Bland &Altman 1983, bias and limits of agreement testing based on summaries per patient
- Percentage time < 10 bpm

So, what did we do?

- For regulators that were concerned with repeated measures:
- Same as for 1)
- Plus: Bootstrapping, one observation per patient, estimate the Mean accuracy and Limits of Agreement and associated Bootstrap confidence limits
- Bland-Altman plots investigating bias vs mean
- Added value of Deming regression not really understood

Result

- First regulatory review resulted in certification
- Awaiting the second regulatory review