PERFORMANCE MEASURES AND EVALUATION

- Estimating performance
- Classic performance metrics
- Performance metrics
- How well are ST episodes detected?
- Characteristics of transient ST episodes
- Estimating performance of ST episode detection
- Estimating performance of detecting total ST episode time
- Estimating performance of measuring the ST segment deviation
- Estimating performance of classifying ischaemic and non-ischaemic heart - rate related ST episodes
- Assessing the robustness (predicting the “real world” performance)

Estimating performance

- Development and testing database ??
- In the field of developing analyzers of biomedical signals use development database and report performance achieved, and possibly try to predict the analyzer's performance in the “real world” (ANSI/AAMI, Us. FDA, Physionet)

  (Examples: MIT/BIH DB, QT DB, ESC DB, LTST DB, TPEHG DB)
- Due to desired comparison of the results of the evaluation, it is necessary to use ALL records of the database available and UNIQUE performance metrics

Classic performance measures

- Performance evaluation

Classic event oriented performance matrix

		Analyzer	Analyzer
		EVENT	NON-EVENT
Reference	event	*TP*	*FN*
Reference	non-event	*FP*	*(TN)*

TP – number of correctly detected events

FN – number of missed events

FP – number of falsely detected events

TN – number of correctly rejected non-events (undefined for detection task !)

Classic performance measures

- Performance evaluation (detection task)

		Analyzer	Analyzer
		EVENT	NON-EVENT
Reference	event	*TP*	*FN*
Reference	non-event	*FP*	*(TN)*

Sensitivity:

$$Se = \frac{TP}{TP + FN}$$

The proportion of EVENTS which were correctly detected as events

Positive predictivity:

$$+P = \frac{TP}{TP + FP}$$

The proportion of detected EVENTS which actually were events

Classic performance measures

- Performance evaluation (classification task)

		Analyzer	Analyzer
		EVENT	NON-EVENT
Reference	event	*TP*	*FN*
Reference	non-event	*FP*	*TN*

Sensitivity:

$$Se = \frac{TP}{TP + FN}$$

The proportion of EVENTS which were correctly classified as events

Positive predictivity:)

$$+P = \frac{TP}{TP + FP}$$

The proportion of classified EVENTS which actually were events

Specificity:

$$Sp = \frac{TN}{TN + FP}$$

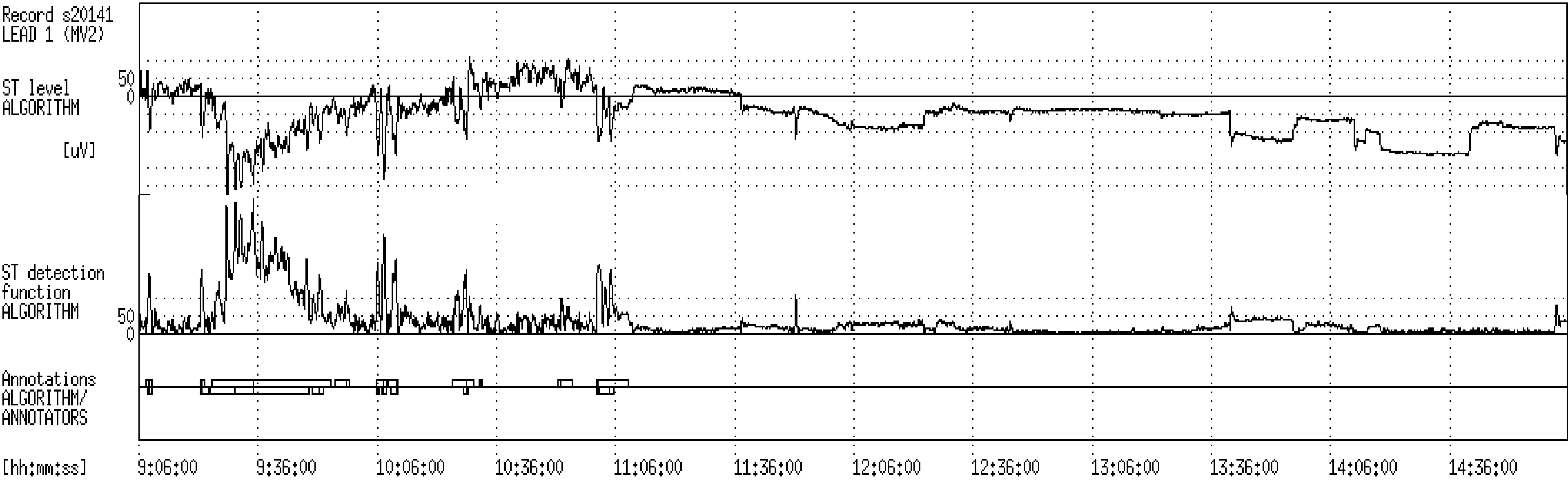
The proportion of NON-EVENTS which were correctly classified as non-events)

Performance metrics

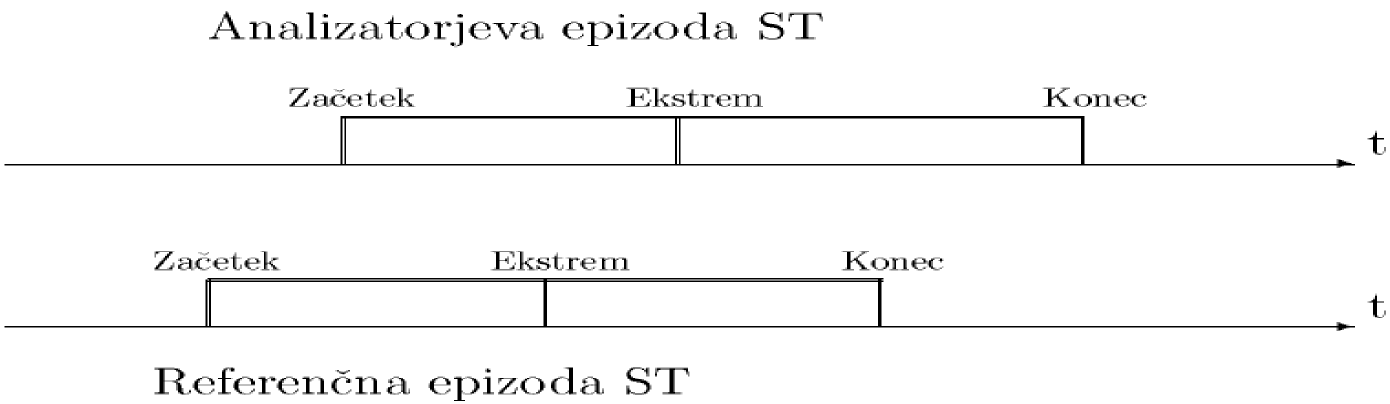
- The evaluation of an ST detection algorithm or analyzer should answer the following questions:

- How well are ST episodes detected ?
- How reliably is ST episode or ischaemic ST episode duration measured ?
- How accurately are ST deviations measured ?
- How well are ischaemic and non-ichaemic heart rate related ST episodes differentiated ?
- How well will the ST analyzer perform in the “real world” ?

How well are ST episodes detected?

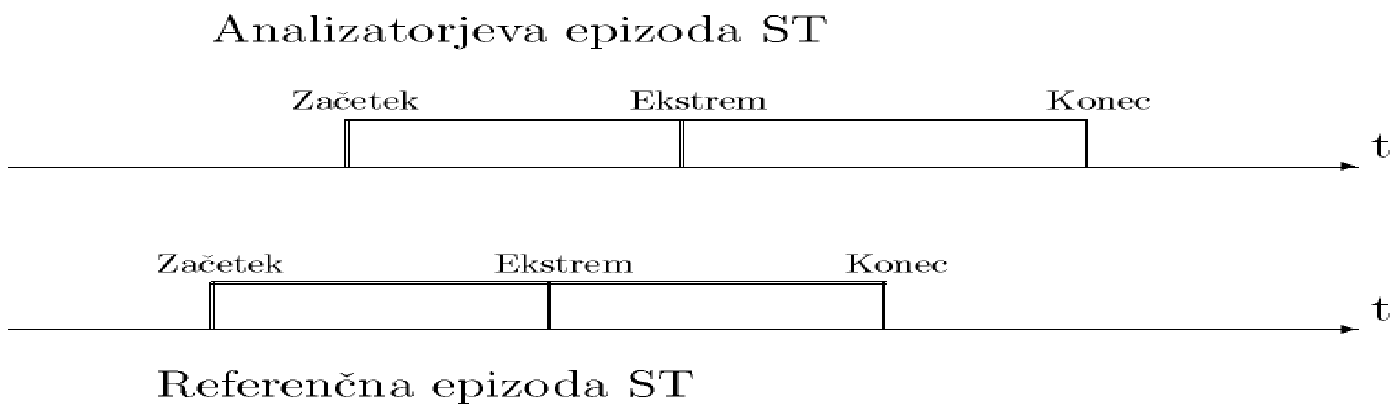


Characteristics of transient ST episodes



- Number
- Type (ischaemic, due to heart frequency changes)
- Length
- ST segment level at extrema deviation

Estimating performance of ST episode detection



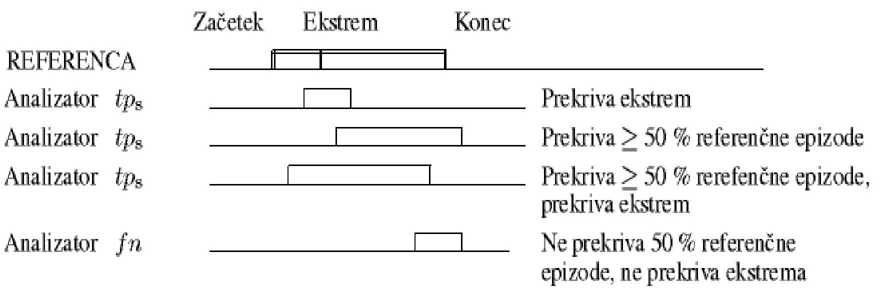
Characteristics of analyzer-annotated and reference episodes:

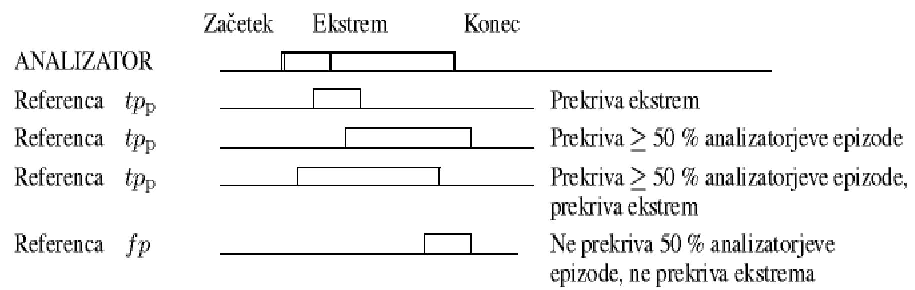
- Episodes contain time dimension
- They differ in length
- There is no one-to-one correspondence
- There is no “non-events”

Estimating performance of ST episode detection

Estimating performance of transient ST segment episode detection

TPs, TPp, FN, FP: Matching test for reference and analyzer-annotated episodes





TPs - Reference episodes that satisfied the matching criteria

TPp - Analyzer-annotated episodes that satisfied the matching criteria

FN - MIssed episodes

FP - Falsely detected episodes

Estimating performance of ST episode detection

Estimating performance of transient ST segment episode detection

Performance measures (ANSI/AAMI, Us. FDA)

Se		Analyzer	
		ST epi.	Other
Ref.	ST epi.	*TPs*	*FN*
	Other	-	-

+P		Analyzer	
		ST epi.	Other
Ref.	ST epi.	*TPp*	-
	Other	*FP*	-

Se - Sensitivity

An estimate of the likelihood of detecting an ST episode

$$Se = TPs \;/\; (TPs + FN)$$

+P – Positive predictivity

An estimate of the likelihood that a detection is a true ST episode:

$$+P = TPp \;/\; (TPp + FP)$$

Estimating performance of ST episode detection

(LTST DB, Protocol B, 908 combined ST segment episodes)
(Time domain and KLT approach)

Se		Algorithm	
		ST epis	Other
Ref	ST epis	723	185
Ref	Other	-	-

+P		Algorithm	
		ST epis	Other
Ref	ST epis	750	-
Ref	Other	208	-

ST epis – Combined ST segment episodes
Se = 79.6%   +P = 78.3%

Estimating performance of ST episode detection

Technique		*ESC DB* *SE* [%] *Se*	*+P*
Time domain	[g]	81	76
	[a]	84	81
RMS method	[g]	–	–
	[a]	84.7	86.1
Time domain	[g]	79.2	81.4
	[a]	81.5	82.5
KLT approach	[g]	85.2	86.2
	[a]	87.1	87.7
Time domain, KLT	[g]	77.2	86.3
	[a]	81.3	89.2
Neural net	[g]	85.0	68.7
	[a]	88.6	78.4
Neural net, KLT	[g]	–	–
	[a]	77	86

Estimating performance of ST episode detection

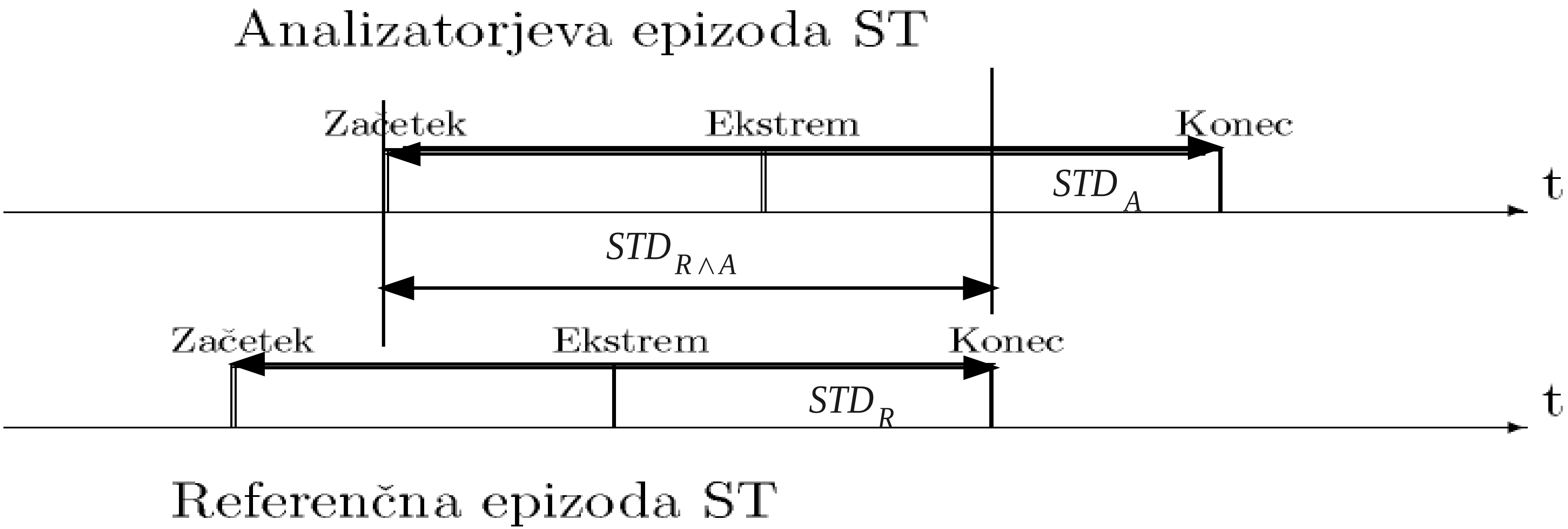
Technique		ESC DB SE [%] Se	ESC DB SE [%] +P		LTST DB SE [%] Se	LTST DB SE [%] +P
Time domain	[g]	81	76		–	–
	[a]	84	81		–	–
RMS method	[g]	–	–		–	–
	[a]	84.7	86.1		–	–
Time domain	[g]	79.2	81.4		–	–
	[a]	81.5	82.5		–	–
KLT approach	[g]	85.2	86.2	–>	77.0	58.8
	[a]	87.1	87.7	–>	74.0	61.4
Time domain, KLT	[g]	77.2	86.3	<–	79.6	78.3
	[a]	81.3	89.2	<–	78.9	80.7
Neural net	[g]	85.0	68.7		–	–
	[a]	88.6	78.4		–	–
Neural net, KLT	[g]	–	–		–	–
	[a]	77	86		–	–

Performance metrics

- The evaluation of an ST detection algorithm or analyzer should answer the following questions:

- How well are ST episodes detected ?
- How reliably is ST episode or ischaemic ST episode duration measured ?
- How accurately are ST deviations measured ?
- How well are ischaemic and non-ichaemic heart rate related ST episodes differentiated ?
- How well will the ST analyzer perform in the “real world” ?

Estimating performance of detecting total ST episode time



$$STD\ Se = \frac{STD_{R \wedge A}}{STD_R} \qquad STD+P = \frac{STD_{R \wedge A}}{STD_A}$$

(ANSI/AAMI, Us. FDA)

Estimating performance of detecting total ST episode time

Technique		ESC DB SE [%] Se	ESC DB SE [%] +P	ESC DB SD [%] Se	ESC DB SD [%] +P		LTST DB (Protocol B) SE [%] Se	LTST DB (Protocol B) SE [%] +P	LTST DB (Protocol B) SD [%] Se	LTST DB (Protocol B) SD [%] +P
Time domain	[g]	81	76	–	–		–	–	–	–
	[a]	84	81	–	–		–	–	–	–
RMS method	[g]	–	–	–	–		–	–	–	–
	[a]	84.7	86.1	75.3	68.2		–	–	–	–
Time domain	[g]	79.2	81.4	–	–		–	–	–	–
	[a]	81.5	82.5	–	–		–	–	–	–
KLT approach	[g]	85.2	86.2	75.8	78.0	–>	77.0	58.8	48.5	47.8
	[a]	87.1	87.7	78.2	74.1	–>	74.0	61.4	54.8	58.4
Time domain, KLT	[g]	77.2	86.3	67.5	69.2	<–	79.6	78.3	68.4	67.3
	[a]	81.3	89.2	77.6	68.9	<–	78.9	80.7	73.1	74.9
Neural net	[g]	85.0	68.7	73.0	69.5		–	–	–	–
	[a]	88.6	78.4	72.2	67.5		–	–	–	–
Neural net, KLT	[g]	–	–	–	–		–	–	–	–
	[a]	77	86	–	–		–	–	–	–

Performance metrics

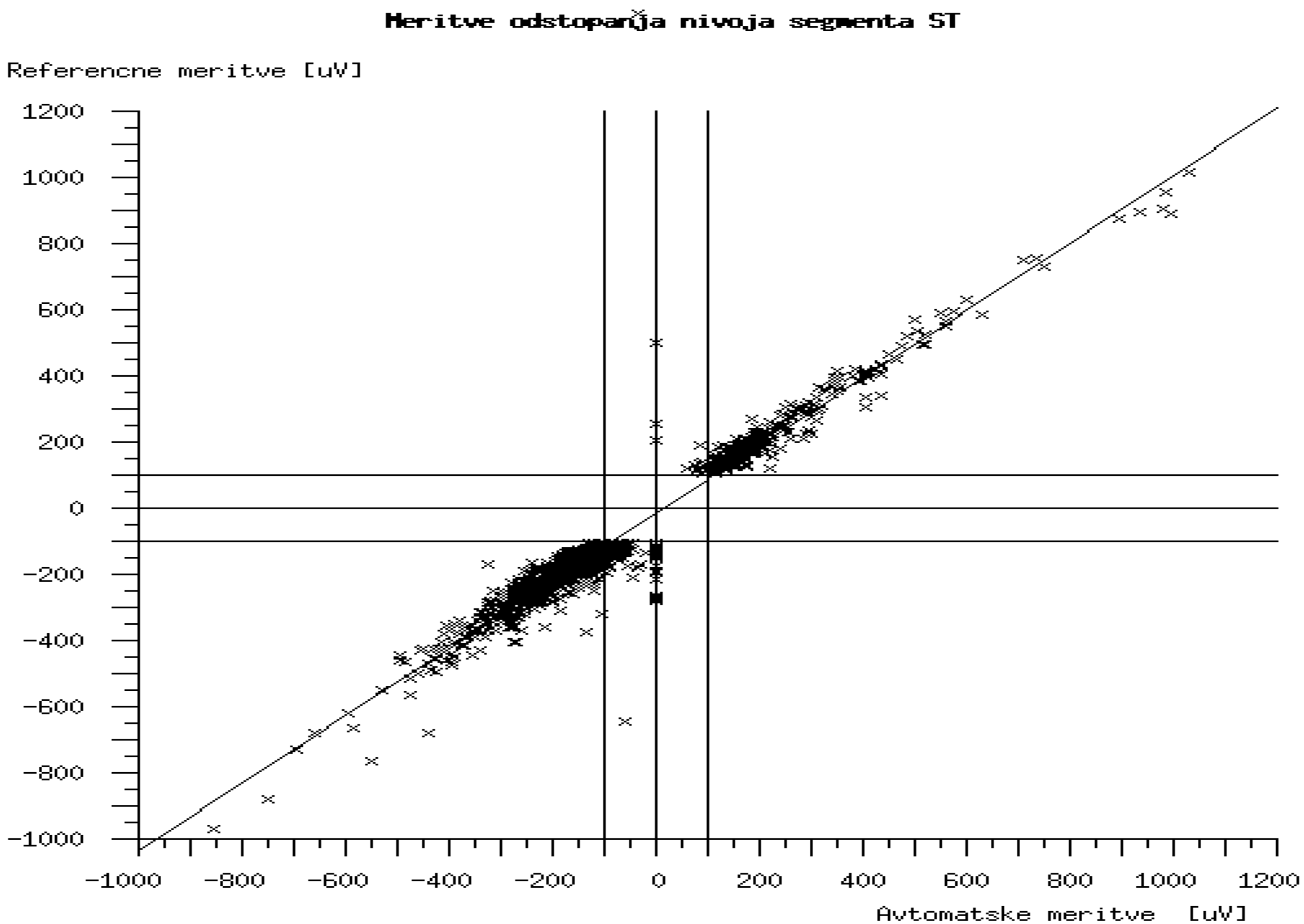
- The evaluation of an ST detection algorithm or analyzer should answer the following questions:

- How well are ST episodes detected ?
- How reliably is ST episode or ischaemic ST episode duration measured ?
- How accurately are ST deviations measured ?
- How well are ischaemic and non-ichaemic heart rate related ST episodes differentiated ?
- How well will the ST analyzer perform in the “real world” ?

Estimating performance of measuring the ST segment deviation

- Measurement error = Measurement of the analyzer - Reference measurement
- Mean error
- Standard deviation
- Correlation coefficient
- Regression line
- Percentage of measurements of which error did not exceed 100µV (Discrepant ST segment measurement percentage, *p* (100 µV))
- Measurement error in µV which was not exceeded by 95% of measurements (Discrepant ST segment deviation measurement value, *e* (95%))

Estimating performance of measuring the ST segment deviation



Stevilo epizod	=	1364	Povprecje [uV]	=	16.41
St. dev. [uV]	=	63.13	Korel. koef.	=	0.96
p(100uV) [%]	=	4.25	e(95%) [uV]	=	90.00
Ref. [uV]	=	1.02 Meritev - 14.36 [uV]			

Performance metrics

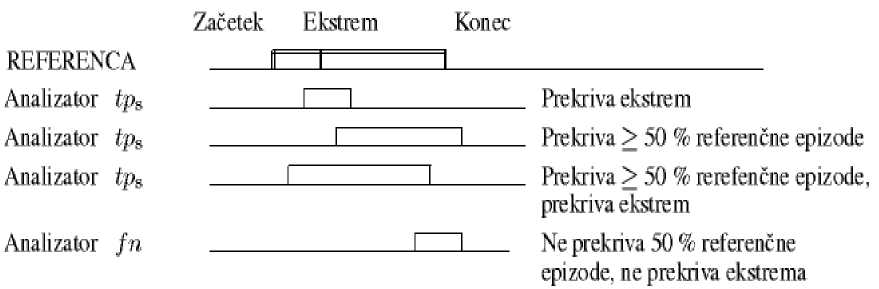
- The evaluation of an ST detection algorithm or analyzer should answer the following questions:

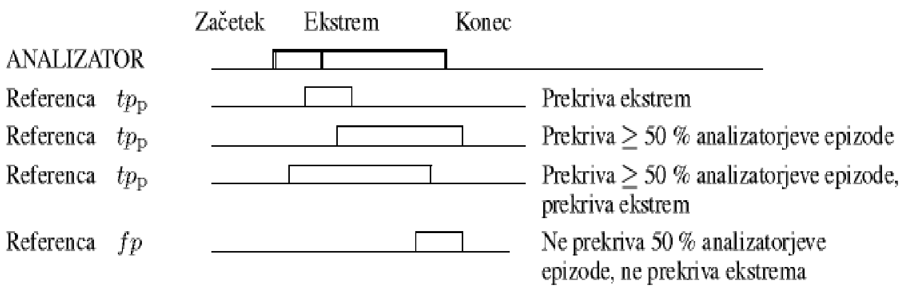
- How well are ST episodes detected ?
- How reliably is ST episode or ischaemic ST episode duration measured ?
- How accurately are ST deviations measured ?
- How well are ischaemic and non-ichaemic heart rate related ST episodes differentiated ?
- How well will the ST analyzer perform in the “real world” ?

Estimating performance of classifying ischaemic and non-ischaemic heart-rate related ST episodes

Estimating performance of transient ST segment episode detection

TPs, *TPp*, *FN*, *FP*: Matching test for reference and analyzer-annotated episodes





TPs - Reference episodes that satisfied the matching criteria

TPp - Analyzer-annotated episodes that satisfied the matching criteria

FN - MIssed episodes

FP - Falsely detected episodes

Estimating performance of classifying ischaemic and non-ischaemic heart-rate related ST episodes

- Extended matching test (*SR(u), SA(v)* – status of reference or analyzer-annotated episode after the matching test)

if *match with ischemic analyzer-annotated ST episodes* **then**
$S_{\mathrm{R}}(u) = \mathit{ischemic};$
else if *match with heart rate related analyzer-annotated ST episodes* **then**
$S_{\mathrm{R}}(u) = \mathit{heart\ rate\ related};$
else
$S_{\mathrm{R}}(u) = \mathit{missed};$
endif

if *match with ischemic reference-annotated ST episodes* **then**
$S_{\mathrm{A}}(v) = \mathit{ischemic};$
else if *match with heart rate related reference-annotated ST episodes* **then**
$S_{\mathrm{A}}(v) = \mathit{heart\ rate\ related};$
else
$S_{\mathrm{A}}(v) = \mathit{falsely\ detected};$
endif

Estimating performance of classifying ischaemic and non-ischaemic heart-rate related ST episodes

Se		Analyzer		
		Isch	Hrr	Other
Ref.	Isch	a	b	c
	Hrr	d	e	f
	Other	-	-	-

+P		Analyzer		
		Isch	Hrr	Other
Ref.	Isch	g	h	-
	Hrr	i	j	-
	Other	k	l	-

Isch - Ischaemic ST segment episodes
Hrr - Heart rate related ST segment episodes
Other - No ST segment changes
ST epi. - ST segment episodes (Isch + Hrr)

Estimating performance of classifying ischaemic and non-ischaemic heart-rate related ST episodes

Se		Analyzer		
		Isch	Hrr	Other
Ref.	Isch	a	b	c
	Hrr	d	e	f
	Other	-	-	-

+P		Analyzer		
		Isch	Hrr	Other
Ref.	Isch	g	h	-
	Hrr	i	j	-
	Other	k	l	-

If considering both ischaemic and heart rate related ST change episodes together as ST change episodes of unique type, then the performance matrices can easily be reduced back to two-by-two, with:

$TP_S = a+b+d+e \qquad FN = c+f$

$TP_P = g+h+i+j \qquad FP = k+l$

$Se = TPs \;/\; (TPs + FN)$

$+P = TPp \;/\; (TPp + FP)$

Estimating performance of classifying ischaemic and non-ischaemic heart-rate related ST episodes

		Analyzer	
		Isch	Hrr
Ref.	Isch	TP	FN
	Hrr	FP	TN

		Analyzer	
		Isch	Hrr
Ref.	Isch	g	h
	Hrr	i	j

Analyzer-annotated ST segment episodes can further be classified between ischaemic and non-ischaemic heart rate related ST episodes. Classic performance measures in terms of *Se* and *Sp* can be used.

$$Se = \frac{TP}{TP + FN} = \frac{g}{g + h}$$

$$Sp = \frac{TN}{TN + FP} = \frac{j}{j + i}$$

Estimating performance of classifying ischaemic and non-ischaemic heart-rate related ST episodes

- Analyzer using time-domain, KLT and Legendre Polynomial Transform approaches

Groups	Gross		Average	
	$Se[\%]$	$Sp[\%]$	$Se[\%]$	$Sp[\%]$
HR, ST	97.5	84.2	95.9	81.8
HR, ST, MD	97.8	85.0	98.4	82.5
HR, LPC	98.5	85.5	97.9	80.4
HR, LPC, MD	$\boxed{98.4}$	$\boxed{85.9}$	$\boxed{98.1}$	$\boxed{85.2}$

Performance metrics

- The evaluation of an ST detection algorithm or analyzer should answer the following questions:

- How well are ST episodes detected ?
- How reliably is ST episode or ischaemic ST episode duration measured ?
- How accurately are ST deviations measured ?
- How well are ischaemic and non-ichaemic heart rate related ST episodes differentiated ?
- How well will the ST analyzer perform in the “real world” ?

Assessing the robustness (predicting the “real world” performance)

- Aggregate gross statistics
- Aggregate average statistics
- “Bootstrap method” of random generating new databases
- Noise stress test (assessing performance after adding noise to records)
- Sensitivity analysis by modifying analyzer's architecture parameters

Assessing the robustness (predicting the “real world” performance)

- Aggregate gross statistics
  - How well an analyzer detects a randomly chosen ST episode ?
- Aggregate average statistics
  - How well the analyzer performs on a randomly chosen ambulatory record ?
- “Bootstrap method” of random generating new databases
- Noise stress test (assessing performance after adding noise to records)
- Sensitivity analysis by modifying analyzer's architecture parameters

Assessing the robustness (predicting the “real world” performance)

Technique		ESC DB SE [%] Se	ESC DB SE [%] +P	ESC DB SD [%] Se	ESC DB SD [%] +P		LTST DB (Protocol B) SE [%] Se	LTST DB (Protocol B) SE [%] +P	LTST DB (Protocol B) SD [%] Se	LTST DB (Protocol B) SD [%] +P
Time domain	[g]	81	76	–	–		–	–	–	–
	[a]	84	81	–	–		–	–	–	–
RMS method	[g]	–	–	–	–		–	–	–	–
	[a]	84.7	86.1	75.3	68.2		–	–	–	–
Time domain	[g]	79.2	81.4	–	–		–	–	–	–
	[a]	81.5	82.5	–	–		–	–	–	–
KLT approach	[g]	85.2	86.2	75.8	78.0	–>	77.0	58.8	48.5	47.8
	[a]	87.1	87.7	78.2	74.1	–>	74.0	61.4	54.8	58.4
Time domain, KLT	[g]	77.2	86.3	67.5	69.2	<–	79.6	78.3	68.4	67.3
	[a]	81.3	89.2	77.6	68.9	<–	78.9	80.7	73.1	74.9
Neural net	[g]	85.0	68.7	73.0	69.5		–	–	–	–
	[a]	88.6	78.4	72.2	67.5		–	–	–	–
Neural net, KLT	[g]	–	–	–	–		–	–	–	–
	[a]	77	86	–	–		–	–	–	–

Assessing the robustness (predicting the “real world” performance)

- Aggregate gross statistics
- Aggregate average statistics
- “Bootstrap method” of random generating new databases
  - Is the analyzer's performance critically dependent on the choice of the test database ?
- Noise stress test (assessing performance after adding noise to records)
- Sensitivity analysis by modifying analyzer's architecture parameters

Assessing the robustness (predicting the “real world” performance)

“Bootstrap” method (necessary assumption: the records of the origin(al) database are representative set of records for the problem domain !):

(1) Randomly (and with replacement) choose *L* elements from the original database and insert them into a hypothetical database

(2) Using the hypothetical database estimate the analyzer's performance

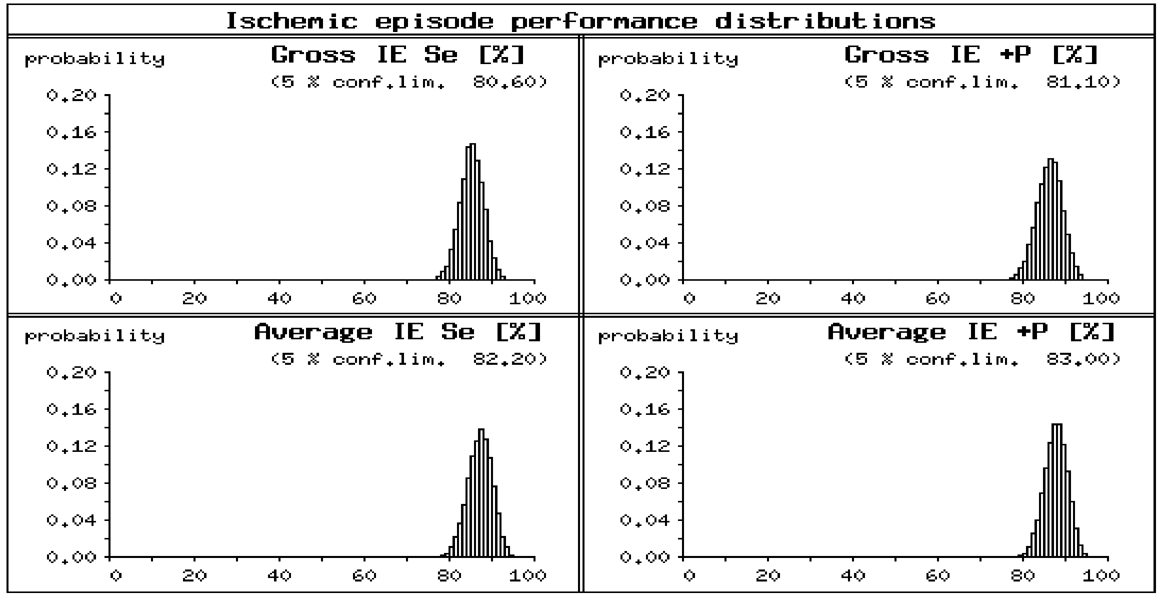
(3) Repeat steps (1) and (2) many (10.000) times

(4) Assess the estimates of the performance *distributions* obtained in the step 2 in the sense of 95% confidence limits

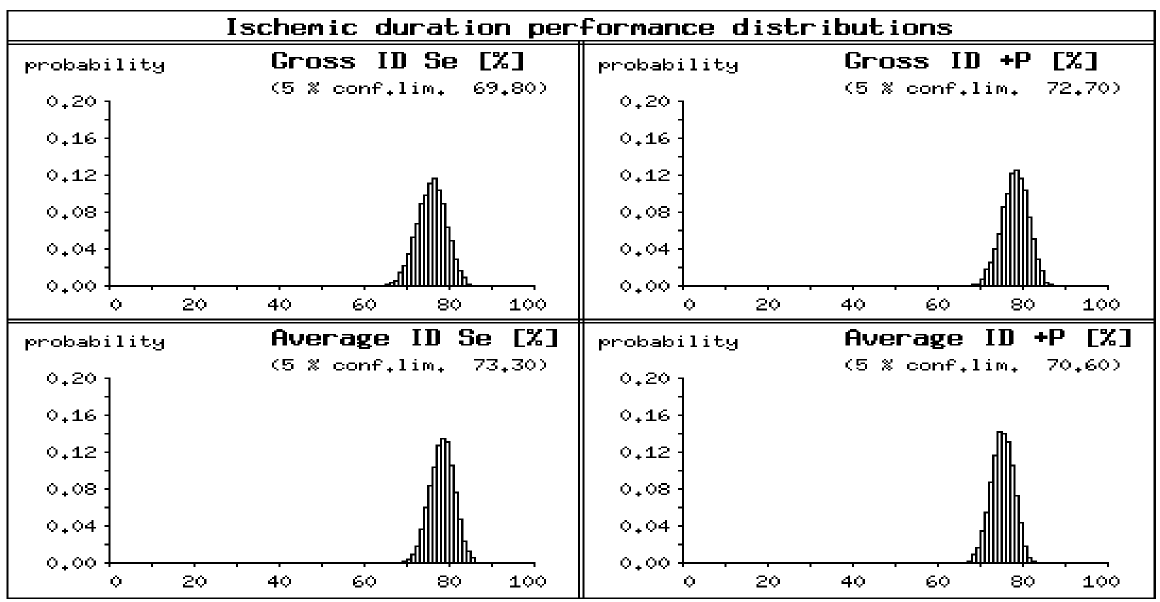
Assessing the robustness (predicting the “real world” performance)

		ESC DB					LTST DB (Protocol B)			
		SE [%]		SD [%]			SE [%]		SD [%]	
Technique		Se	+P	Se	+P		Se	+P	Se	+P
Time domain	[g]	81	76	–	–		–	–	–	–
	[a]	84	81	–	–		–	–	–	–
RMS method	[g]	–	–	–	–		–	–	–	–
	[a]	84.7	86.1	75.3	68.2		–	–	–	–
Time domain	[g]	79.2	81.4	–	–		–	–	–	–
	[a]	81.5	82.5	–	–		–	–	–	–
KLT approach	[g]	85.2	86.2	75.8	78.0	–>	77.0	58.8	48.5	47.8
	[a]	87.1	87.7	78.2	74.1	–>	74.0	61.4	54.8	58.4
Time domain, KLT	[g]	77.2	86.3	67.5	69.2	<–	79.6	78.3	68.4	67.3
	[a]	81.3	89.2	77.6	68.9	<–	78.9	80.7	73.1	74.9
Neural net	[g]	85.0	68.7	73.0	69.5		–	–	–	–
	[a]	88.6	78.4	72.2	67.5		–	–	–	–
Neural net, KLT	[g]	–	–	–	–		–	–	–	–
	[a]	77	86	–	–		–	–	–	–

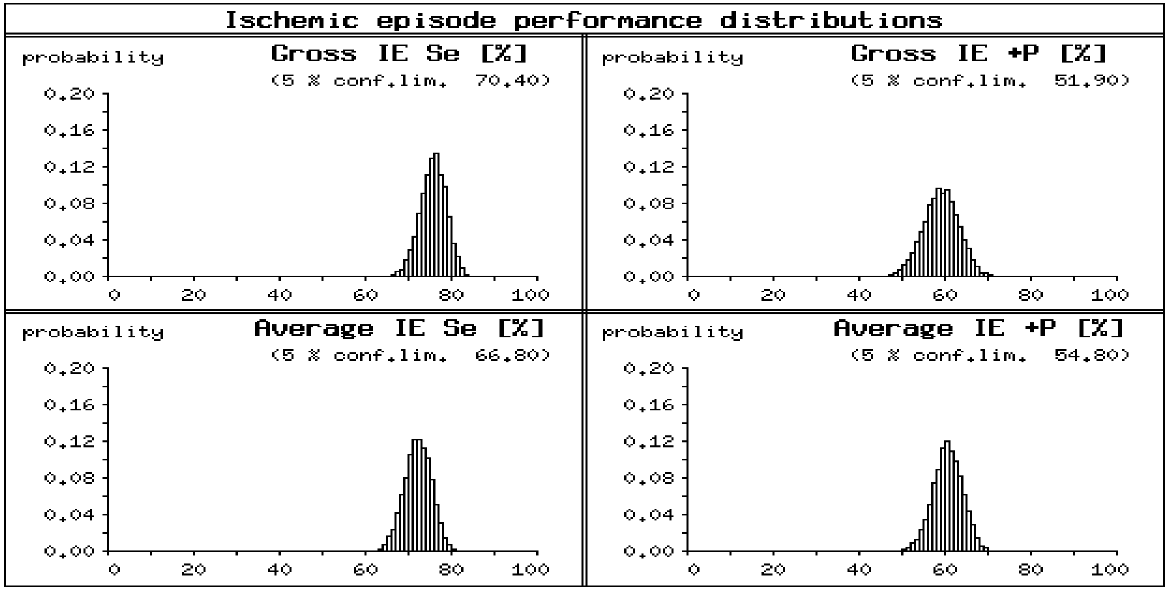
Assessing the robustness (predicting the "real world" performance)

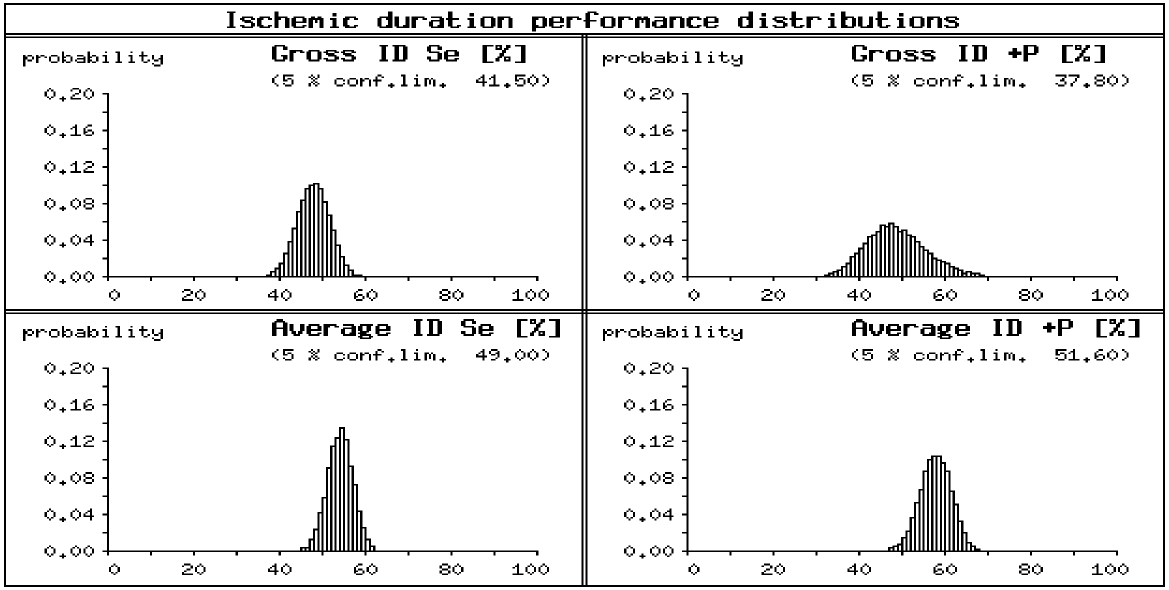


- KLT approach, "bootstrap" distributions as obtained on ESC DB (development database)



Assessing the robustness (predicting the “real world” performance)

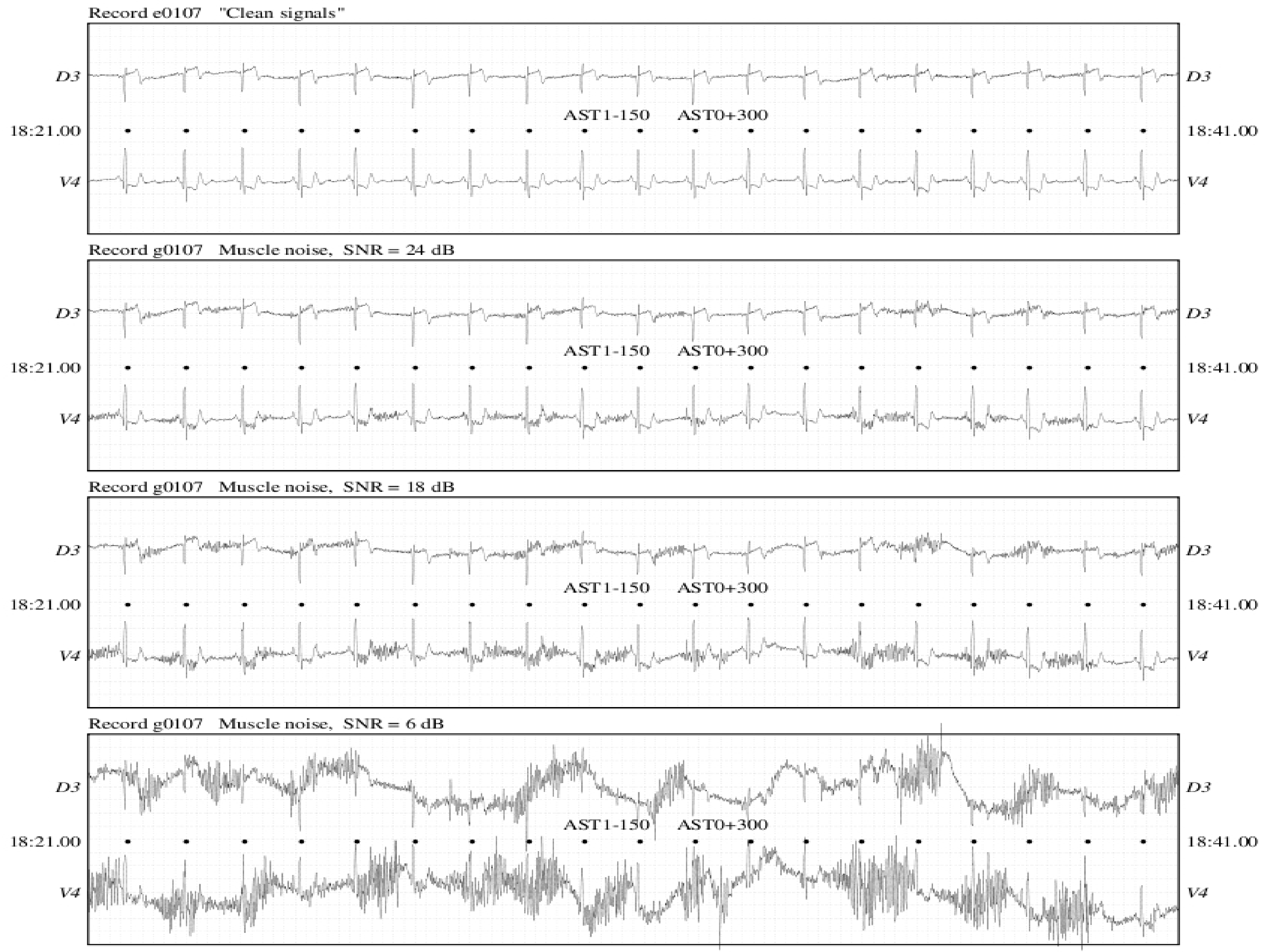




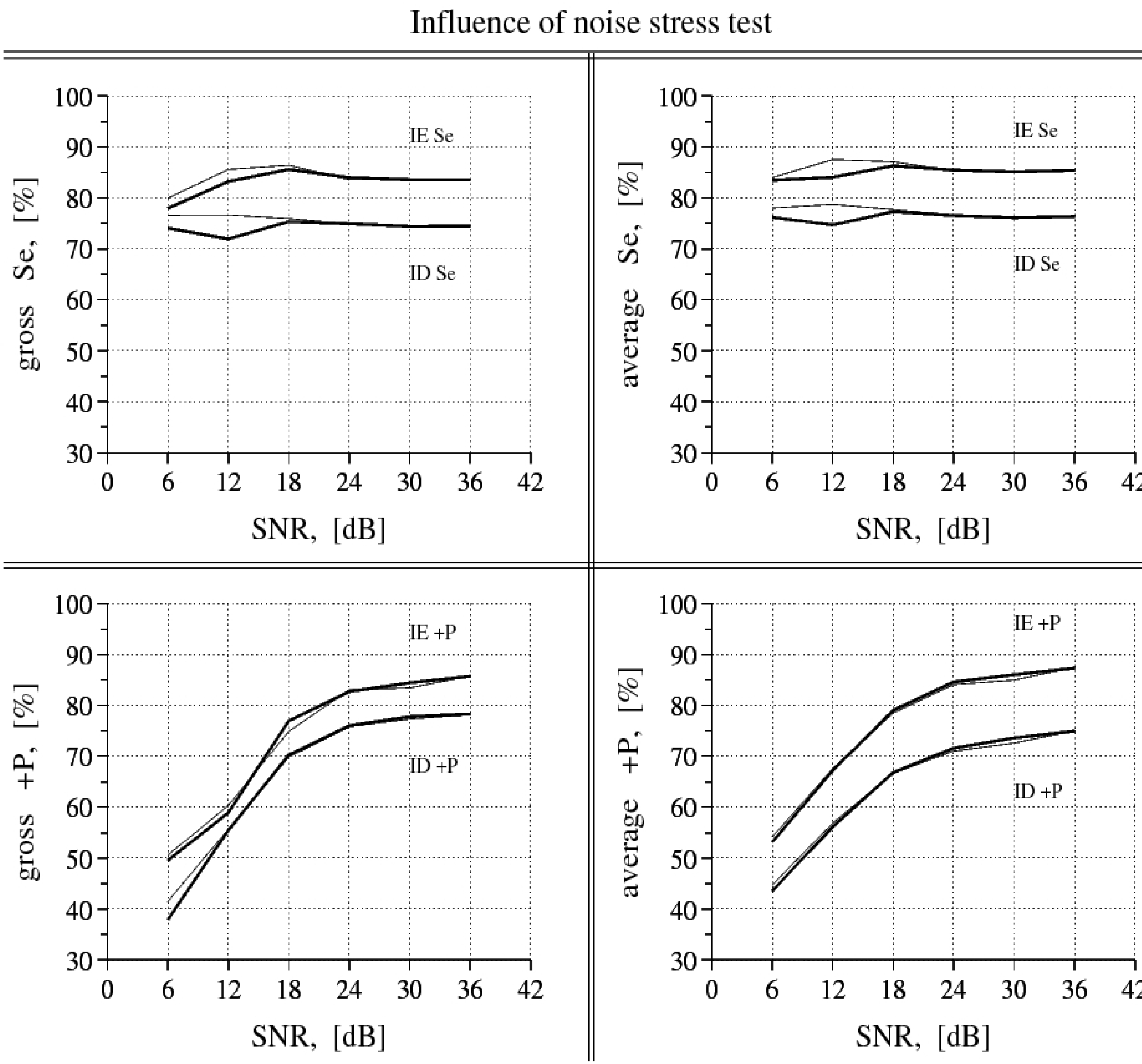
- KLT approach, “bootstrap” distributions as obtained on LTST DB (testing database)

Assessing the robustness (predicting the “real world” performance)

- Aggregate gross statistics
- Aggregate average statistics
- “Bootstrap method” of random generating new databases
- Noise stress test (assessing performance after adding noise to records)
  - What is the minimum critical signal-to-noise ratio at which the analyzer's performance is still acceptable ?
- Sensitivity analysis by modifying analyzer's architecture parameters


Record e0107 "Clean signals"
D3
18:21.00
AST1-150
AST0+300
18:41.00
V4
Record g0107 Muscle noise, SNR = 24 dB
D3
18:21.00
AST1-150
AST0+300
18:41.00
V4
Record g0107 Muscle noise, SNR = 18 dB
D3
18:21.00
AST1-150
AST0+300
18:41.00
V4
Record g0107 Muscle noise, SNR = 6 dB
D3
18:21.00
AST1-150
AST0+300
18:41.00
V4

Assessing the robustness (predicting the “real world” performance)



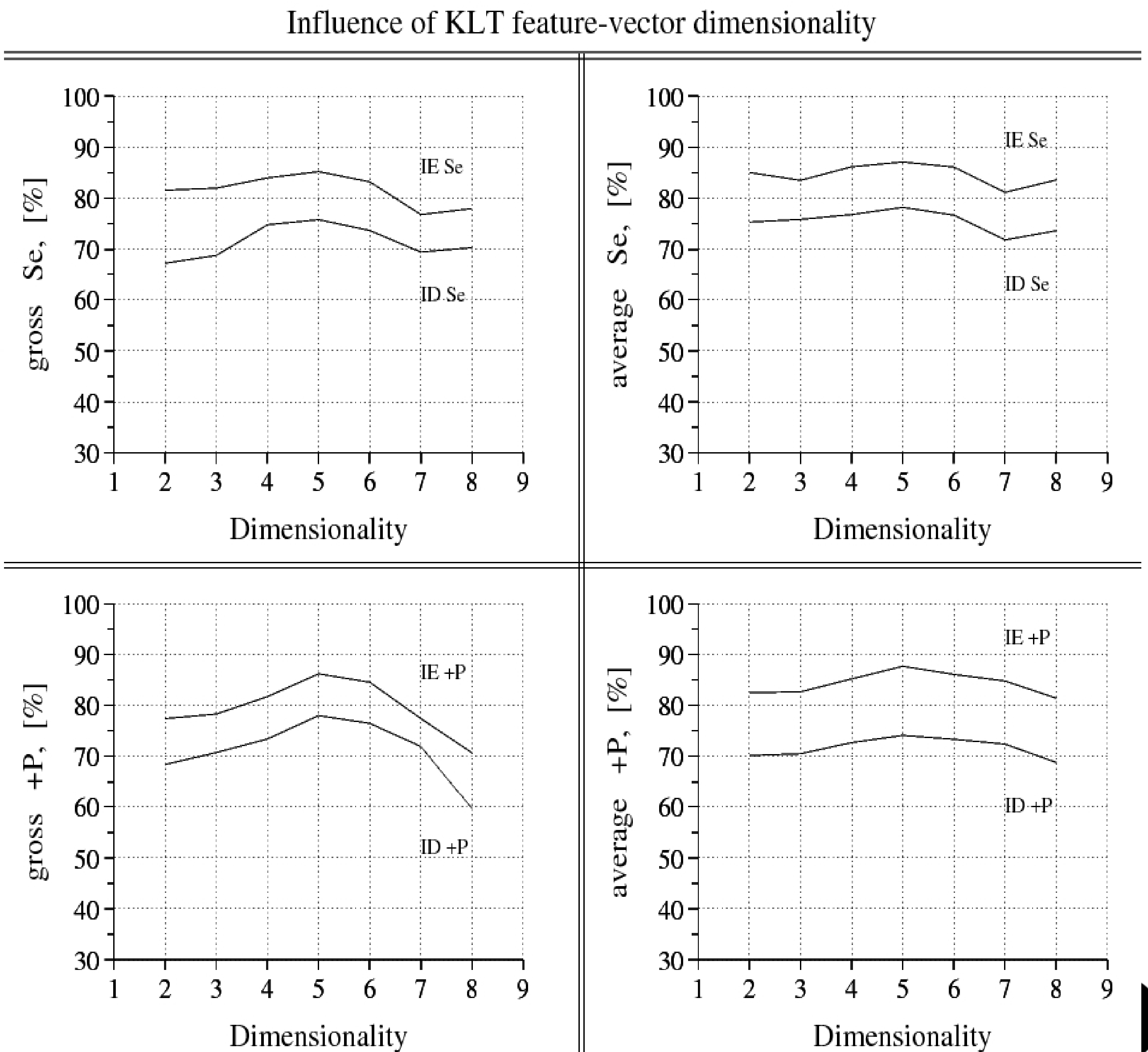
- KLT approach, “bootstrap” distributions as obtained on ESC DB (development database)

Assessing the robustness (predicting the "real world" performance)

- Aggregate gross statistics
- Aggregate average statistics
- "Bootstrap method" of random generating new databases
- Noise stress test (assessing performance after adding noise to records)
- Sensitivity analysis by modifying analyzer's architecture parameters

  \- Are the architecture parameters critically tuned to the development database ?

Assessing the robustness (predicting the “real world” performance)



- KLT approach, “bootstrap” distributions as obtained on ESC DB (development database)