Clinical Data Miner (CDM)

Towards more efficient clinical study support

Arnaud Installé
arnaud.installe@esat.kuleuven.be

19 June 2014

Outline

- Clinical diagnostic model research
- Electronic Data Capture (EDC)
- Integrated data analysis
- Machine-learning automation
- Conclusions & Future work

Clinical diagnostic model research

Why?

- Importance of diagnostic procedures:
  - Diagnosis $\rightarrow$ disease management
  - UK, 2009: late diagnosis $\rightarrow$ 5k - 10k cancer deaths

Clinical diagnostic model research

Why?

- Importance of diagnostic procedures:
  - Diagnosis $\rightarrow$ disease management
  - UK, 2009: late diagnosis $\rightarrow$ 5k - 10k cancer deaths
- Impediments:
  - Cost
  - Invasive
  - False positives

Clinical diagnostic model research

Why?

- Importance of diagnostic procedures:
  - Diagnosis $\rightarrow$ disease management
  - UK, 2009: late diagnosis $\rightarrow$ 5k - 10k cancer deaths
- Impediments:
  - Cost
  - Invasive
  - False positives

$\Rightarrow$ Clinical diagnostic model research

Clinical diagnostic model research

How?

Patient data

Clinical diagnostic model research

How?

Patient data

$+$

Gold standard diagnosis

Clinical diagnostic model research

How?

Patient data

$+$

Gold standard diagnosis

$\downarrow$

Machine-learning model

Clinical diagnostic model research

How?

Patient data

$+$

Gold standard diagnosis

$\downarrow$

Machine-learning model

$\downarrow$

Diagnostic predictions

Clinical diagnostic model research

International Ovarian Tumour Analysis (IOTA)

- Started in 2000 by Timmerman, Bourne, Valentin
- Consortium of gynaecologists specialized in ultrasound
- Aim: diagnostic models for ovarian tumours

Clinical diagnostic model research

International Ovarian Tumour Analysis (IOTA)

- Started in 2000 by Timmerman, Bourne, Valentin
- Consortium of gynaecologists specialized in ultrasound
- Aim: diagnostic models for ovarian tumours

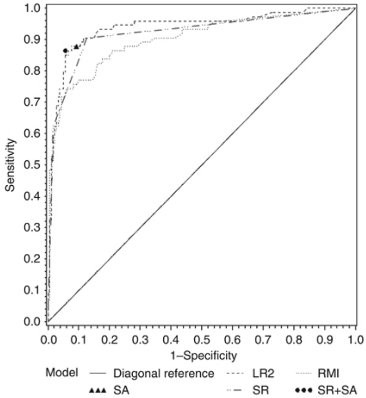
⇒ Several models a.o. LR2 (Timmerman et al., 2005)

Clinical diagnostic model research

International Ovarian Tumour Analysis (IOTA)

- Started in 2000 by Timmerman, Bourne, Valentin
- Consortium of gynaecologists specialized in ultrasound
- Aim: diagnostic models for ovarian tumours

⇒ Several models a.o. LR2 (Timmerman et al., 2005)



ROC curves for several ovarian tumour diagnostic models, a.o. RMI and LR2. Source: Sayasneh et al. (2013) .

Clinical diagnostic model research

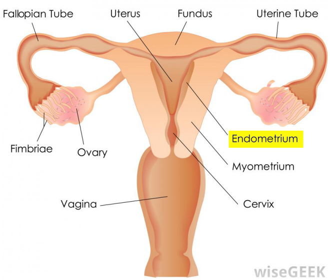
International Endometrial Tumour Analysis (IETA)

- Started in 2008 by Van den Bosch, Timmerman
- Consortium of gynaecologists specialized in ultrasound imaging

Clinical diagnostic model research

International Endometrial Tumour Analysis (IETA)

- Started in 2008 by Van den Bosch, Timmerman
- Consortium of gynaecologists specialized in ultrasound imaging
- Aim: diagnostic models for endometrial pathology (a.o. cancer)

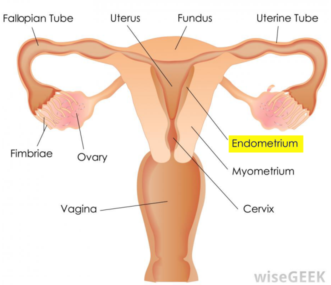


Source: http://www.wisegeek.org/what-is-the-endometrial-cavity.htm

Clinical diagnostic model research

International Endometrial Tumour Analysis (IETA)

- Started in 2008 by Van den Bosch, Timmerman
- Consortium of gynaecologists specialized in ultrasound imaging
- Aim: diagnostic models for endometrial pathology (a.o. cancer)
- 3 different studies; overlapping CRFs: common variables: 75 (nominal: 31; ordinal: 8; continuous: 14; text: 13; date: 4)

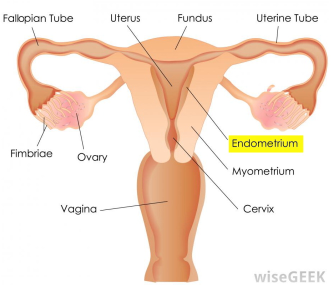


Source: http://www.wisegeek.org/what-is-the-endometrial-cavity.htm

Clinical diagnostic model research

International Endometrial Tumour Analysis (IETA)

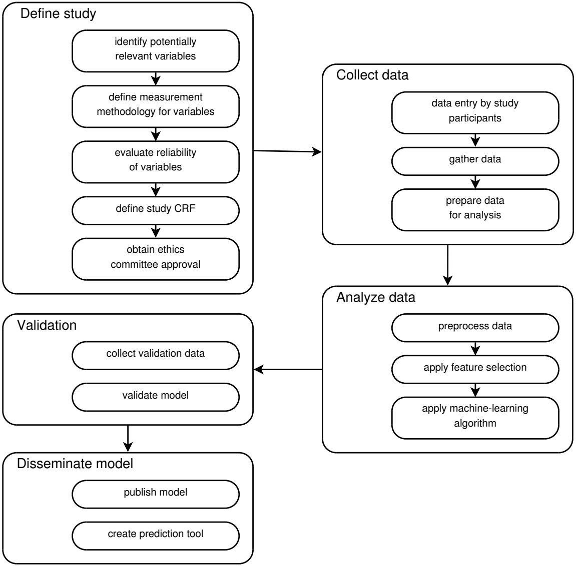
- Started in 2008 by Van den Bosch, Timmerman
- Consortium of gynaecologists specialized in ultrasound imaging
- Aim: diagnostic models for endometrial pathology (a.o. cancer)
- 3 different studies; overlapping CRFs: common variables: 75 (nominal: 31; ordinal: 8; continuous: 14; text: 13; date: 4)
- Similar workflow as for IOTA studies



Source: http://www.wisegeek.org/what-is-the-endometrial-cavity.htm

Clinical diagnostic model research

Workflow



Clinical diagnostic model research

International Endometrial Tumour Analysis (IETA)

Ultrasound Obstet Gynecol 2010; **35**: 103–112
Published online 15 December 2009 in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/uog.7487

Terms, definitions and measurements to describe the sonographic features of the endometrium and intrauterine lesions: a consensus opinion from the International Endometrial Tumor Analysis (IETA) group

F. P. G. LEONE*, D. TIMMERMAN†, T. BOURNE‡, L. VALENTIN§, E. EPSTEIN¶, S. R. GOLDSTEIN**, H. MARRET††, A. K. PARSONS‡‡, B. GULL§§, O. ISTRE¶¶, W. SEPULVEDA***, E. FERRAZZI††† and T. VAN DEN BOSCH†

*Departments of Obstetrics and Gynecology, *Clinical Sciences Institute L. Sacco and †††Children's Hospital, Buzzi, University of Milan, Milan, Italy, †Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Leuven, Belgium, ‡Department of Obstetrics and Gynaecology, Imperial College London, London, UK, Departments of Obstetrics and Gynecology, §Malmö University Hospital and ¶Lund University Hospital, Lund University, Lund and §§University of Göteborg, Sahlgrenska University Hospital, Göteborg, Sweden, Departments of Obstetrics and Gynecology, **New York University School of Medicine, New York, NY, ‡‡University of South Florida, Tampa, FL and ¶¶Harvard Medical School, Boston, MA, USA, ††Service de Gynécologie, Centre Hospitalier Universitaire Bretonneau, Tours, France and ***Department of Obstetrics and Gynecology, San Jose Hospital, University of Santiago de Chile, Santiago, Chile*

KEYWORDS: diagnosis; endometrial disease; endometrial neoplasms; endometrium; myoma; sonography; terminology; ultrasonography; uterus

ABSTRACT

The IETA (International Endometrial Tumor Analysis group) statement is a consensus statement on terms,

with an endometrial thickness of 5 mm or more, an evaluation of endometrial morphology and vascularization using gray-scale and Doppler ultrasound imaging with or without the added use of sonohysterography (instillation

IETA consensus paper

Clinical diagnostic model research

International Endometrial Tumour Analysis (IETA)

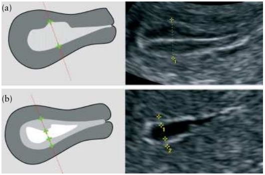


Figure 2 (a) Diagram and accompanying ultrasound image showing measurement of the endometrial thickness in the absence of intracavitary fluid; the endometrium should be measured where it appears to be at its thickest. (b) When intracavitary fluid is present, the thickness of both single layers is measured in the sagittal plane and the sum is recorded. The measurement should be taken where the endometrium appears to be at its thickest. If the endometrium is thickened asymmetrically, the anterior and posterior endometrial thicknesses should also be reported separately.

diameters in millimeters, rounded up to one decimal point. The volume of the lesion may be calculated from the three orthogonal diameters using the formula for a prolate ellipsoid ($d1 \times d2 \times d3 \times 0.523$). In myomas, the distance from the back of the myoma to the serosa should also be measured if a surgical resection is considered.

The amount of intracavitary fluid is defined by its symmetrical anterior and posterior sides. This definition includes the different appearances seen throughout the menstrual cycle and the monolayer pattern found in most postmenopausal patients. A 'uniform' endometrium includes the three-layer pattern, as well as the homogeneous hyperechogenic, hypoechogenic and isoechogenic endometrium (Figure 3). The echogenicity is defined as 'non-uniform' if the endometrium appears heterogeneous, asymmetrical or cystic (Figure 4).

The endometrial midline is defined as 'linear', if a straight hyperechogenic interface within the endometrium is visualized, as 'non-linear' if a waved hyperechogenic interface is seen, and as 'irregular' or as 'not defined' in the absence of a distinct interface (Figure 5).

The 'bright edge' is the echo formed by the interface between an intracavitary lesion and the endometrium (Figure 6)[16].

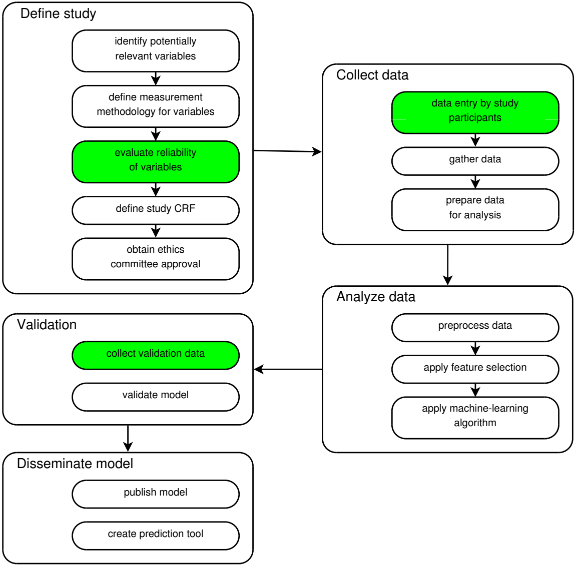
In some patients the endometrial interface is better detected by gently pushing the transvaginal probe against the uterine corpus, which makes the two endometrial surfaces slide against each other (i.e. the 'sliding sign'). This technique may also be used to help characterize pathology, as small amounts of fluid in the cavity may help delineate structures in the cavity.

The endometrial–myometrial junction[17] should be described as 'regular', 'irregular', 'interrupted' or 'not defined' (Figure 7).

Synechiae are defined as strands of tissue crossing the endometrium (Figure 8). Congenital anomalies are not

IETA consensus paper

Electronic Data Capture



Electronic Data Capture

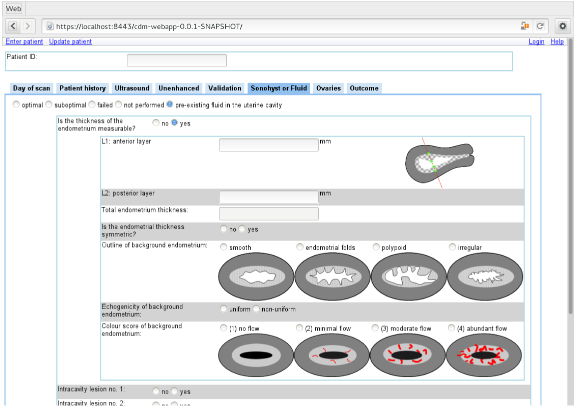
Requirements

- Inclusion of pictograms
- Multi-centre
- Userfriendly
- Generic

Electronic Data Capture

Requirements

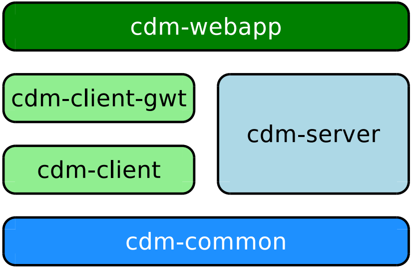
- Inclusion of pictograms
- Multi-centre
- Userfriendly
- Generic



CDM's data collection user interface

Electronic Data Capture

Methodology



Modular, layered architecture

Electronic Data Capture

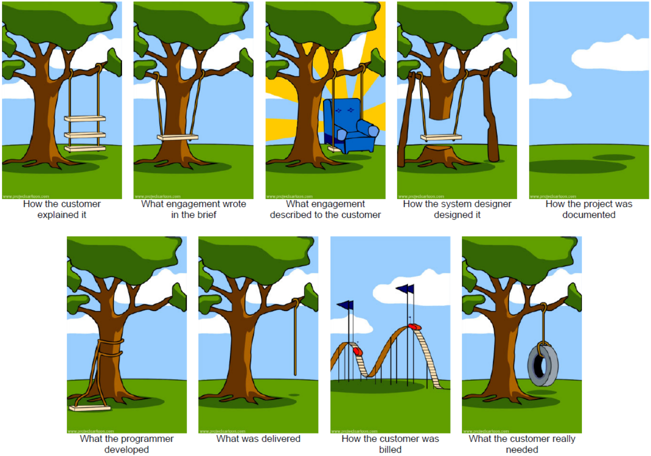
Methodology

- Scrum
  - Regular meetings
  - End-user input

Electronic Data Capture

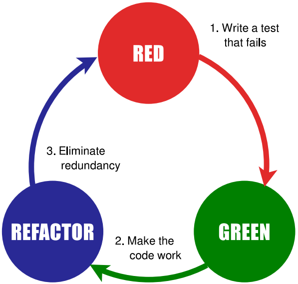
Methodology

- Scrum
  - Regular meetings
  - End-user input



Electronic Data Capture

Methodology



- Test-Driven Development (TDD)
  - Good test coverage guarantees quality
- Design patterns
- Dependency injection

→ Maintainable

Electronic Data Capture

Methodology

Module (SLOC)	Production code (SLOC)	Test code (SLOC)	Line coverage (%)	Branch coverage (%)
cdm-common	5862	7023	91	94
cdm-server	15260	28109	92	90
cdm-client	3595	7607	88	91
cdm-client-gwt	4090	5123	53	42
cdm-webapp	321	177	34	100
	(sum)		(weighted average)	
Overall	29128	48039	85	84

Source Lines of Code (SLOC) and test coverage per module.

Electronic Data Capture

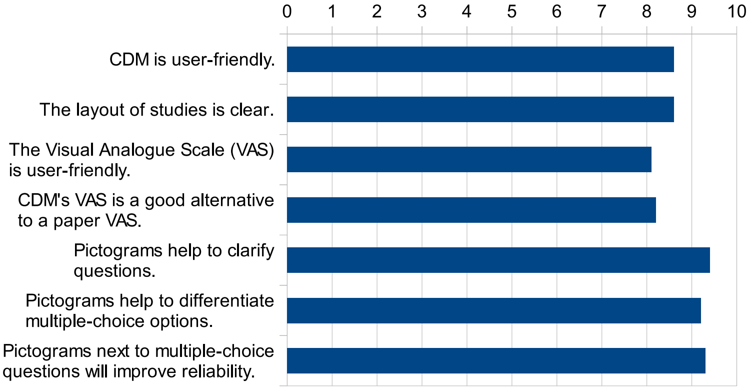
Survey

- Sent to 42 experienced CDM users
  - $>$ 10 IETA entries, or
  - inter-rater agreement study

$\Rightarrow$ 28 responses
(66.7% response rate)

Electronic Data Capture

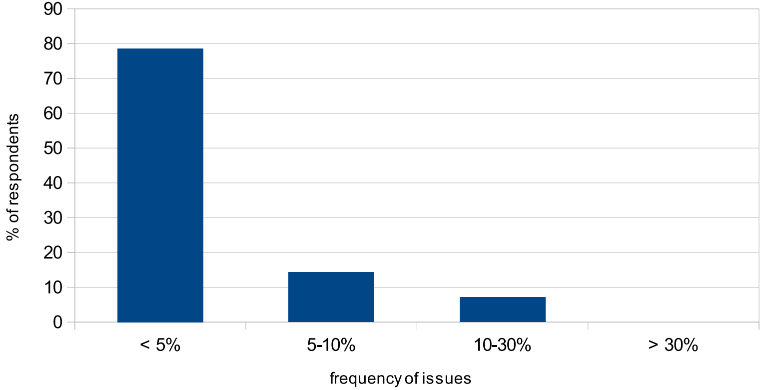
Survey



Average agreement level (0 = no agreement; 10 = full agreement).

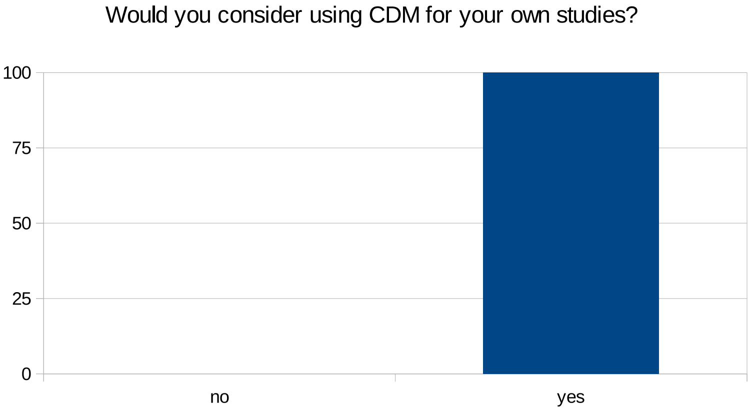
Electronic Data Capture

Survey



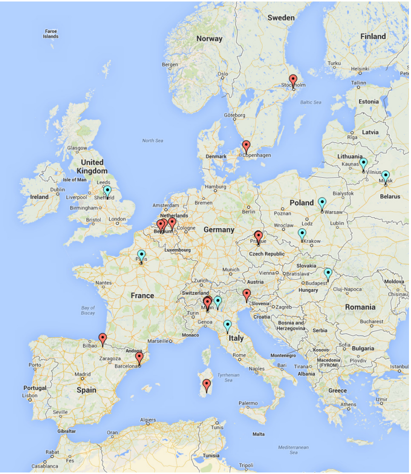
Percentage of CDM user interactions exhibiting software issues.

Electronic Data Capture
Survey



Electronic Data Capture

IETA Data collection



24 centres; 39 participants

Electronic Data Capture

IETA Data collection

Study	Complete	Incomplete	Total
IETA #1a	603	142	745
IETA #1b	723	192	915
IETA #1c	165	96	261
IETA #3	562	120	682
IETA #4	667	423	1090
Total	2720	973	3693

Inclusion numbers

Electronic Data Capture

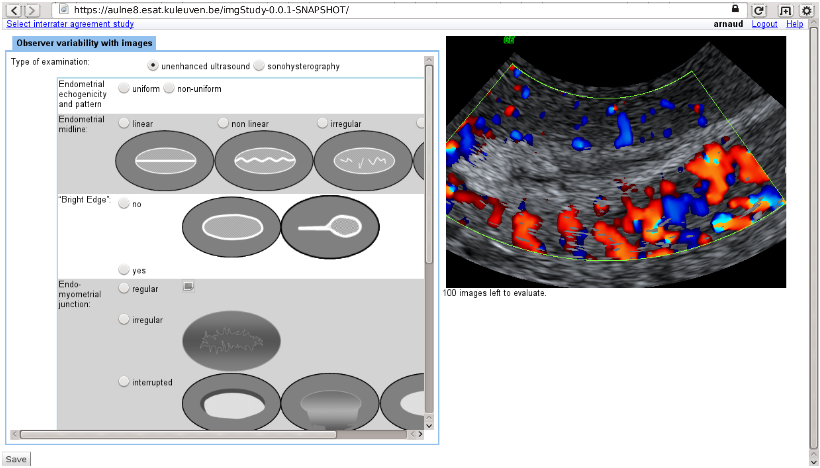
Studies

IETA #1, #3, #4	Van den Bosch, Timmerman
cytoreduction study	Testa
IPULA	Bourne, Condous
IETA #5	Van den Bosch, Fenning

Ongoing and future CDM studies.

Electronic Data Capture

User interface for inter-rater agreement studies



Modified CDM user interface

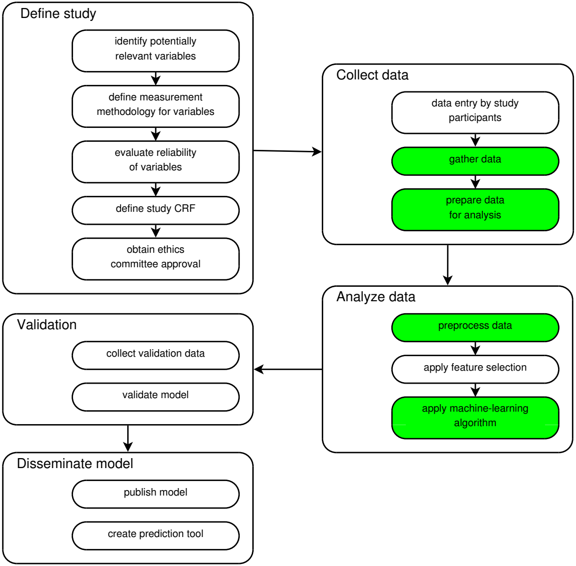
Electronic Data Capture

User interface for inter-rater agreement studies

Influence of pictograms	Installé
Polycystic ovaries	Van Schoubroeck
Uterine anomalies	Van Schoubroeck
Endomyometrial junction	Votino
IETA #2	Valentin
Image enhancement	Bourne

Inter-rater agreement studies.

Integrated data analysis



Integrated data analysis

Advantages

- Export/import
- CRF structure

Integrated data analysis

Data querying & preprocessing

DataManager

```
+newDescriptor(studyId:String): DataDescriptor
+stripText(descriptor:DataDescriptor): DataDescriptor
+stripOptional(descriptor:DataDescriptor): DataDescriptor
+stripDates(descriptor:DataDescriptor): DataDescriptor
+flatten(descriptor:DataDescriptor): DataDescriptor
+createFactorProxies(descriptor:DataDescriptor): DataDescriptor
+normalize(descriptor:DataDescriptor): DataDescriptor
+select(descriptor:DataDescriptor,
        identifiers:String[]): DataDescriptor
+deselect(descriptor:DataDescriptor,
          identifiers:String[]): DataDescriptor
+merge(descriptors:DataDescriptor[]): DataDescriptor
+label(descriptor:DataDescriptor,
       identifiers:String[]): LabelledDataDescriptor
+load(descriptor:DataDescriptor): Data
+shuffle(data:LabelledData,random:Random): LabelledData
+stratify(data:LabelledData,outputField:Field,
          relativeSizes:int[]): LabelledData[]
+subset(data:LabelledData,fromIndex:int,
        toIndex:int): LabelledData
+dump(data:Data,writer:PrintWriter)
```

- Dummy variables
- Merge studies
- Structurally missing variables

Integrated data analysis

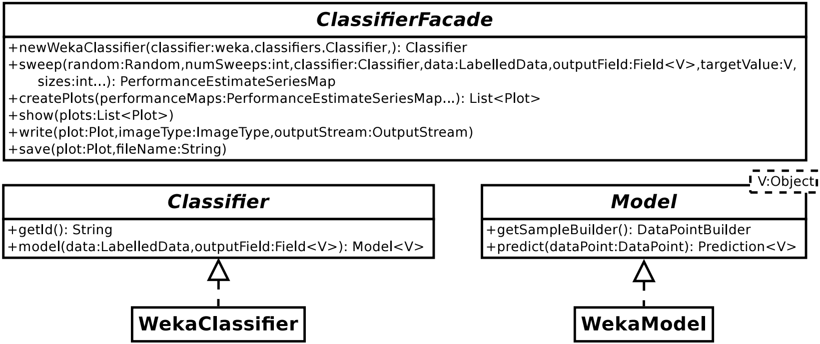
Data querying & preprocessing

```
DataManager
+newDescriptor(studyId:String): DataDescriptor
+stripText(descriptor:DataDescriptor): DataDescriptor
+stripOptional(descriptor:DataDescriptor): DataDescriptor
+stripDates(descriptor:DataDescriptor): DataDescriptor
+flatten(descriptor:DataDescriptor): DataDescriptor
+createFactorProxies(descriptor:DataDescriptor): DataDescriptor
+normalize(descriptor:DataDescriptor): DataDescriptor
+select(descriptor:DataDescriptor,
        identifiers:String[]): DataDescriptor
+deselect(descriptor:DataDescriptor,
          identifiers:String[]): DataDescriptor
+merge(descriptors:DataDescriptor[]): DataDescriptor
+label(descriptor:DataDescriptor,
       identifiers:String[]): LabelledDataDescriptor
+load(descriptor:DataDescriptor): Data
+shuffle(data:LabelledData,random:Random): LabelledData
+stratify(data:LabelledData,outputField:Field,
          relativeSizes:int[]): LabelledData[]
+subset(data:LabelledData,fromIndex:int,
        toIndex:int): LabelledData
+dump(data:Data,writer:PrintWriter)
```

- Dummy variables
- Merge studies
- Structurally missing variables
- Randomize order
- Stratify

Integrated data analysis

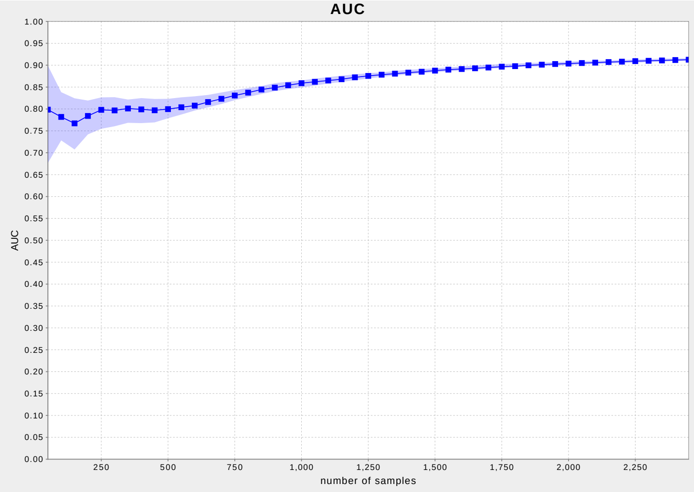
Machine-learning



- (Currently:) classifiers → WEKA
- Prediction outcomes: binary / probability
- Learning curves

Integrated data analysis

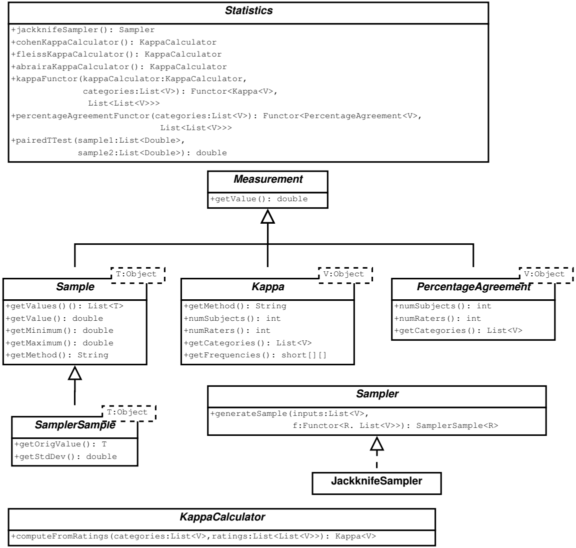
Machine-learning



CDM's machine-learning API enables generation of learning curves.

Integrated data analysis

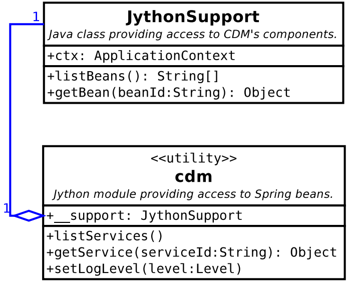
Statistical analysis



- inter-rater agreement: Cohen, Fleiss
- percentage agreement
- jackknife sampling
- paired t-test

Integrated data analysis

Interactive scripting



- bridge Jython $\leftrightarrow$ CDM

Integrated data analysis

Interactive scripting

<<utility>>

dm

Exposes CDM's DataManager interface

```
+__dm: DataManager
```

```
+study(studyId:String): DataDescriptor
+acceptFilter(): DataPointFilter
+rejectFilter(): DataPointFilter
+duplicatesFilter(): DataPointFilter
+suffixFilter(field:Field,s:String): DataPointFilter
+equalsFilter(field:Field,s:String): DataPointFilter
+inFilter(field:Field,ary:String[]): DataPointFilter
+notFilter(filter:DataPointFilter): DataPointFilter
+andFilter(f1:DataPointFilter,f2:DataPointFilter): DataPointFilter
+stripOptional(desc:DataDescriptor): DataDescriptor
+stripText(desc:DataDescriptor): DataDescriptor
+stripDates(desc:DataDescriptor): DataDescriptor
+select(desc:DataDescriptor,identifiers:String[]): DataDescriptor
+deselect(desc:DataDescriptor,identifiers:String[]): DataDescriptor
+label(desc:DataDescriptor,identifiers:String[]): DataDescriptor
+addPatientId(desc:DataDescriptor): DataDescriptor
+addUserName(desc:DataDescriptor): DataDescriptor
+createFactorProxies(desc:DataDescriptor): DataDescriptor
+flatten(desc:DataDescriptor): DataDescriptor
+normalize(desc:DataDescriptor): DataDescriptor
+filter(desc:DataDescriptor,filter:DataPointFilter): DataDescriptor
+load(desc:DataDescriptor): Data
+shuffle(data:Data): Data
+subset(data:Data,fromIndex:int,toIndex:int): Data
+stratify(data:Data,outputField:Field,relativeSizes:int[]): Data[]
+dumpFields(fields:Field[])
+findField(fields:Field[],fieldPath:String): Field
```

- bridge Jython $\leftrightarrow$ CDM
- data querying / preprocessing

Integrated data analysis

Interactive scripting

```
<<utility>>
ml
-----------------------------------------------------------------------
+__classifierFacade: ClassifierFacade
-----------------------------------------------------------------------
+newWekaClassifier(classifier:weka.classifiers.Classifier): Classifier
+getClassifierMap()
+sweepClassifierPerformances(classifiers:Classifier[],
                             data:Data,outputField:Field<V>,
                             targetValue:V,
                             numSweeps:int,
                             sizes:int[]): PerformanceEstimateSeriesMap[]
+createPlots(performanceMaps:PerformanceEstimateSeriesMaps[]): Plot[]
+showPlots(plots:Plot[])
```

- bridge Jython ↔ CDM
- data querying / preprocessing
- machine-learning

Integrated data analysis

Interactive scripting

```
<<utility>>
stats
-----------------------------------------------
+__stat: Statistics
-----------------------------------------------
+mean(sample:Double[]): Double
+stddev(sample:Double[]): Double
+median(sample:Double[]): Double
+percentiles(sample:Double[],percentages:int[]): Double
+jackkifeSampler(): Sampler
+cohenKappaCalculator(): KappaCalculator
+fleissKappaCalculator(): KappaCalculator
+kappaFunctor(calculator:KappaCalculator,
              categories:List<V>): Functor
+percentageAgreementFunctor(categories:List<V>): KappaFunctor
+pairedTTest(sample1:Double[],sample2:Double[]): double
```

- bridge Jython $\leftrightarrow$ CDM
- data querying / preprocessing
- machine-learning
- statistical analysis

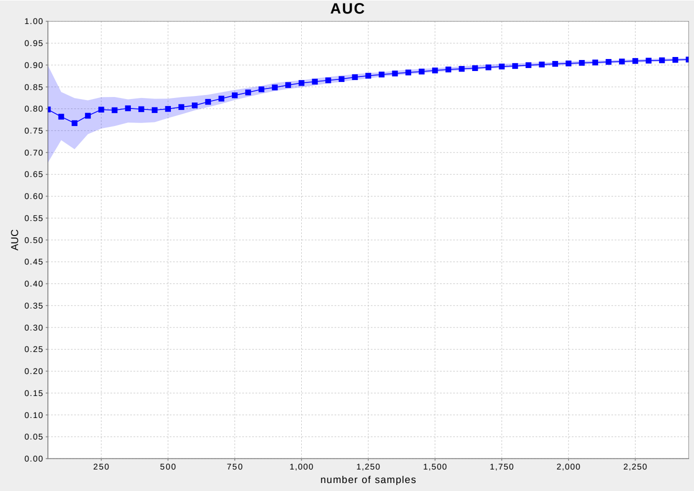
Integrated data analysis

Example 1: Learning curve

```
import dm

# 1. Specify data to load
study_ids = [ 'ieta_1a', 'ieta_1b', 'ieta_1c',
              'ieta_3', 'ieta_4' ]

desc_1 = dm.merge(*[ dm.study(id) for id in study_ids ])

# 2. Preprocess data
desc_2 = dm.stripOptional(dm.stripText(dm.stripDates(desc_1)))
desc_3 = dm.label(desc_2, 'ieta_outcome.*')
desc_4 = dm.createFactorProxies(dm.flatten(dm.normalize(desc_3)))
data = desc_4.load()

# 3. Calculation of learning curves
import weka.classifiers.functions.Logistic as Logistic
import ml

classifier = ml.newWekaClassifier(Logistic())
output_field = dm.findField(data.getOutputFields(),
    'ieta_outcome.endometrium.malignancy')
perf_map = ml.sweep(classifier, data, output_field,
    'yes', 50, range(50, data.size(), 50))

# 4. Generate and save plots
plots = ml.create_plots([ perf_map ], [ 'IETA' ])

for plot in plots:
    ml.savePlot(plot, plot.getTitle() + '.pdf')
```

Integrated data analysis

Example 1: Learning curve



Learning curve generated by script (outcome variable: endometrial malignancy)

Integrated data analysis

Example 2: Contingency table

```
import copy
import dm

# 1. Load combined data of all IETA studies and select fields
study_ids = [ 'ieta_1a', 'ieta_1b', 'ieta_1c', 'ieta_3', 'ieta_4' ]
data = dm.merge(*[ dm.study(id) for id in study_ids ]).load()
field1 = dm.findField(data.getFields(), 'ieta_hist.menopausal_status')
field2 = dm.findField(data.getFields(), 'ieta_outcome.endometrium')

# 3. Count occurrence of each category.
inner = dict([ (c, 0) for c in field1.type().getValues() ])
outer = dict([ (c, copy.copy(inner)) for c in field2.type().getValues() ])

for p in data:
    v1 = p.getValue(field1)
    v2 = p.getValue(field2)

    if v2 != None:
        outer[v2][v1] += 1
    else:
        outer['N/A'][v1] += 1

# 4. Print distribution
total = float(data.size())
print 'outcome,', ', '.join([ d for d in field1.type().getValues() ])

print
for c in field2.type().getValues():
    print c,
    for d in field2.type().getValues():
        print ', %.1f' % (outer[c][d] * 100 / total),
    print
```

Integrated data analysis

Example 2: Contingency table

Outcome	Pre-menopausal (%)	Post-menopausal (%)
atrophy	1.3	9.1
proliferative endometrium	8.0	1.9
secretory endometrium	7.0	0.6
hyperplasia without atypia	2.8	1.3
atypical hyperplasia	0.3	0.3
malignancy	2.6	22.9
endometrial polyp	14.1	11.5
intracavitary myoma	4.1	0.6
endometritis	0.3	0.0
other	1.8	1.1
N/A	5.7	2.6

Contingency table tabulating the frequency distribution of menopausal status versus outcome.

Machine-learning automation

Problem

- Can machine-learning be automated?
  - Data analysis APIs integrated
  - IOTA study $\rightarrow$ required manual preprocessing

Machine-learning automation

Methodology

- Example: IOTA data
  - 46 variables
  - 3511 data points

Machine-learning automation

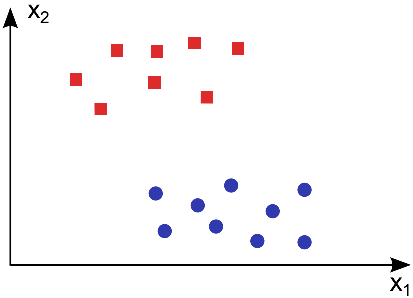
Methodology

- Example: IOTA data
  - 46 variables
  - 3511 data points
- Machine-learning algorithms
  - Logistic regression
  - Least-Squares Support Vector Machines (LS-SVM)

Machine-learning automation

Methodology

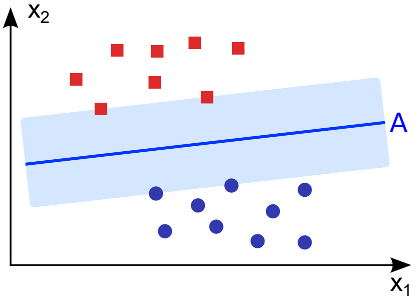
- Example: IOTA data
  - 46 variables
  - 3511 data points
- Machine-learning algorithms
  - Logistic regression
  - Least-Squares Support Vector Machines (LS-SVM)



Machine-learning automation

Methodology

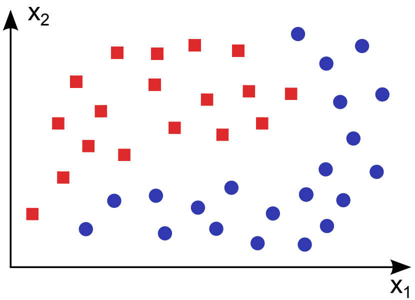
- Example: IOTA data
  - 46 variables
  - 3511 data points
- Machine-learning algorithms
  - Logistic regression
  - Least-Squares Support Vector Machines (LS-SVM)



Machine-learning automation

Methodology

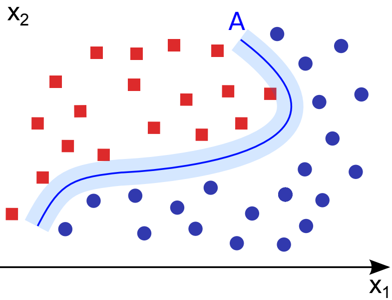
- Example: IOTA data
  - 46 variables
  - 3511 data points
- Machine-learning algorithms
  - Logistic regression
  - Least-Squares Support Vector Machines (LS-SVM) + RBF kernel



Machine-learning automation

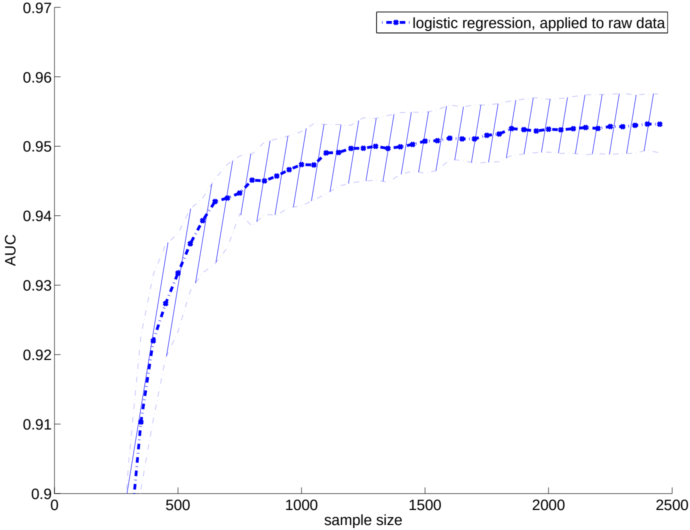
Methodology

- Example: IOTA data
  - 46 variables
  - 3511 data points
- Machine-learning algorithms
  - Logistic regression
  - Least-Squares Support Vector Machines (LS-SVM) + RBF kernel



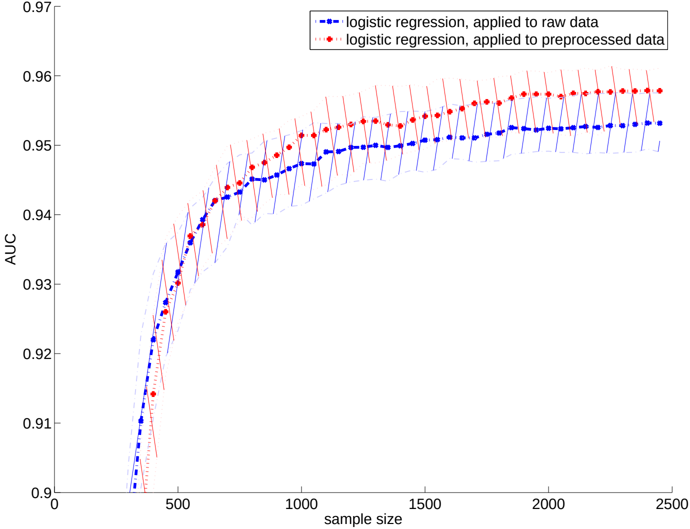
Machine-learning automation

Results



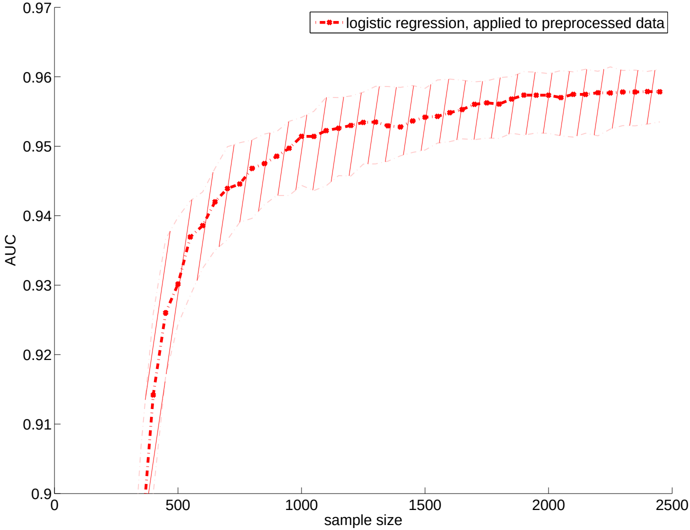
Machine-learning automation

Results



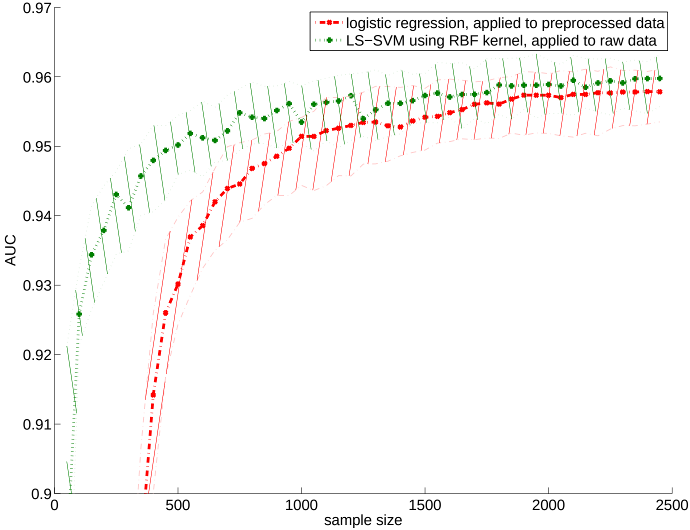
Machine-learning automation

Results



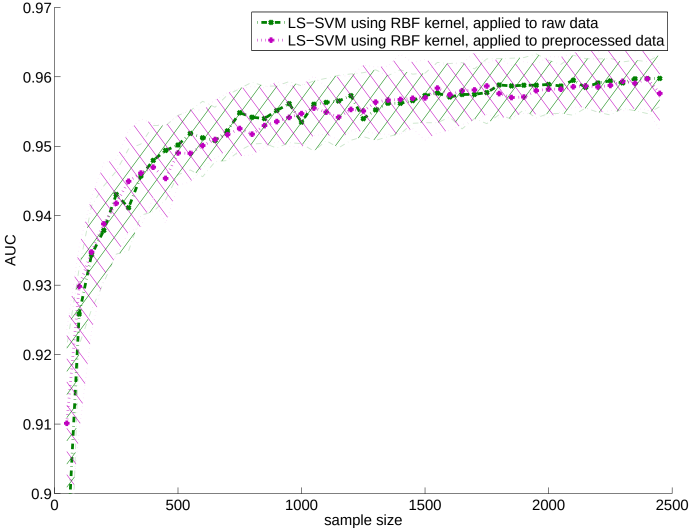
Machine-learning automation

Results



Machine-learning automation

Results



Machine-learning automation

Conclusion

- No need for complex, manual preprocessing
  - Considerably simplifies automation

Machine-learning automation

Conclusion

- No need for complex, manual preprocessing
  - Considerably simplifies automation
- Limitations:
  - Analysis on single data set
  - No interpretability
  - No feature selection in workflow

$\rightarrow$ Future work

Conclusions & future work

Conclusions & future work

Conclusions

- UI for patient data collection
  - User-friendly
  - Multi-centric
  - Generic
  - High user satisfaction

Conclusions & future work

Conclusions

- UI for patient data collection
  - User-friendly
  - Multi-centric
  - Generic
  - High user satisfaction
- UI for inter-rater agreement studies

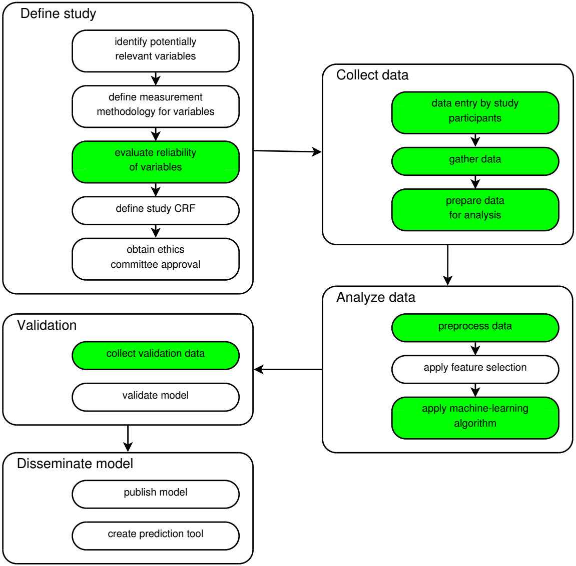
Conclusions & future work

Conclusions

- UI for patient data collection
  - User-friendly
  - Multi-centric
  - Generic
  - High user satisfaction
- UI for inter-rater agreement studies
- Data analysis integration
  - No data export & import
  - Preprocessing
  - Learning curves

Conclusions & future work

Conclusions



Conclusions & future work

Future work

- Goals of TBM: "Endometrial cancer diagnosis based on predictive computer models within an International Endometrial Tumour Analysis (IETA) collaboration"
  - IETA diagnostic model
  - Integrate model in prediction UI

Conclusions & future work

Future work

- Goals of TBM: "Endometrial cancer diagnosis based on predictive computer models within an International Endometrial Tumour Analysis (IETA) collaboration"
  - IETA diagnostic model
  - Integrate model in prediction UI
- Initiate other studies

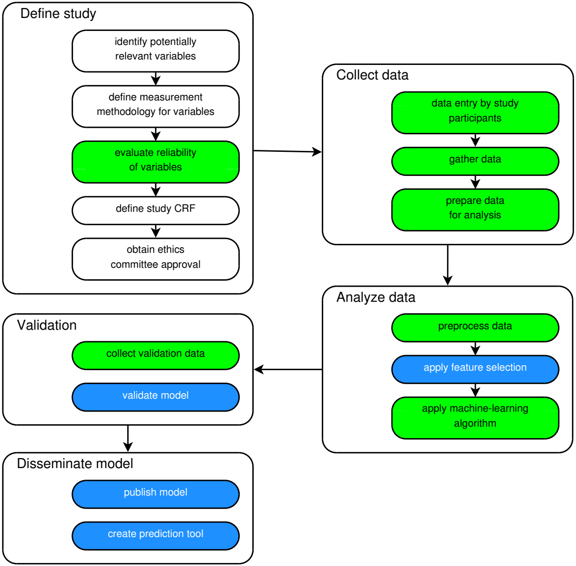
Conclusions & future work

Future work

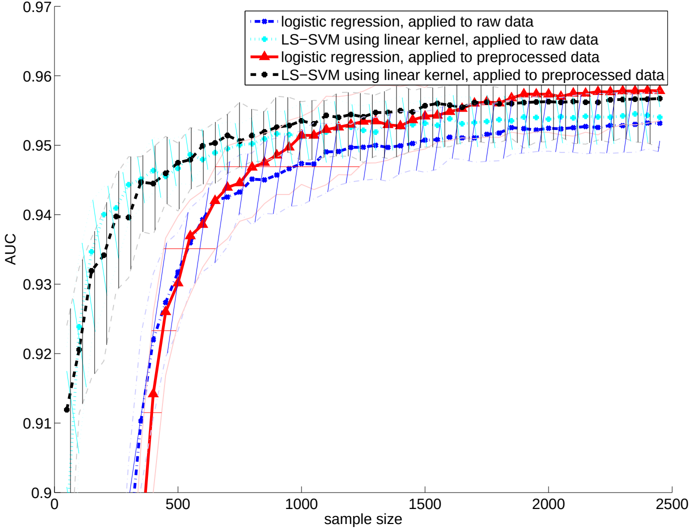
- Goals of TBM: “Endometrial cancer diagnosis based on predictive computer models within an International Endometrial Tumour Analysis (IETA) collaboration”
  - IETA diagnostic model
  - Integrate model in prediction UI
- Initiate other studies
- CDM
  - More sophisticated preprocessors & M-L algorithms
  - UI for study coordinators
  - Teaching tool

Conclusions & future work

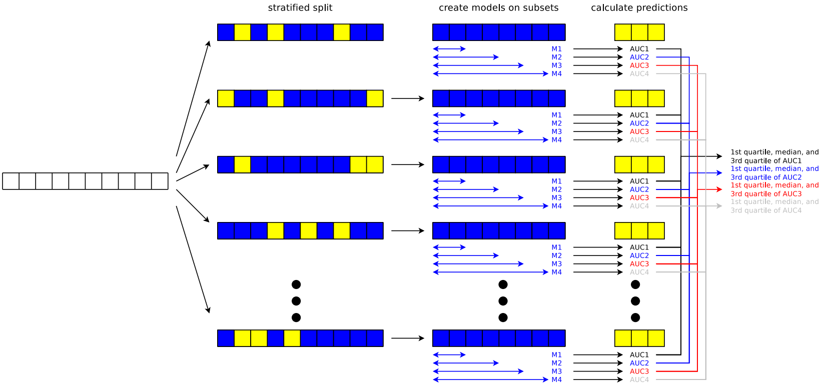
Future work



Thank you!



logistic regression, applied to raw data
LS−SVM using linear kernel, applied to raw data
logistic regression, applied to preprocessed data
LS−SVM using linear kernel, applied to preprocessed data
AUC
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
0
500
1000
1500
2000
2500
sample size


stratified split
create models on subsets
calculate predictions
M1
M2
M3
M4
AUC1
AUC2
AUC3
AUC4
1st quartile, median, and
3rd quartile of AUC1
1st quartile, median, and
3rd quartile of AUC2
1st quartile, median, and
3rd quartile of AUC3
1st quartile, median, and
3rd quartile of AUC4