Belgian Cancer Registry

Stichting tegen Kanker

Fondation contre le Cancer

Semi-automated extraction of information from pathology reports: proof of concept

Joris Mattheijssens, Antoine Pironet, Kris Henau, Petra Denolf, Nancy Van Damme, Harlinde De Schutter

GRELL congress, Thursday May 25th, 2017

www.kankerregister.org I www.registreducancer.org

Objectives

Belgian Cancer Registry (BCR) data collection

- Oncology care programs
  - all new cancer incidences
  - since 2004
- Pathology reports
  - breast, colon and cervical specimens
  - since 2008

Some data are structured (tabulated)

- Patient identification data, incidence date, topography, morphology,…

Some data are unstructured (texts)

- -> extract detailed information
- -> (semi-)automatically

2 case studies

- KRAS mutation in colorectal cancer
- Results of HPV tests in cervical smears

Case study 1 – KRAS in colorectal cancer

Introduction

Ca. 9,000 colorectal cancer incidences/year

KRAS mutation status has influence on choice of therapy

Text collection

- 11,446 pathology reports (C18-19-20, primary tumors)
  containing the search term "KRAS"
  2004-2014

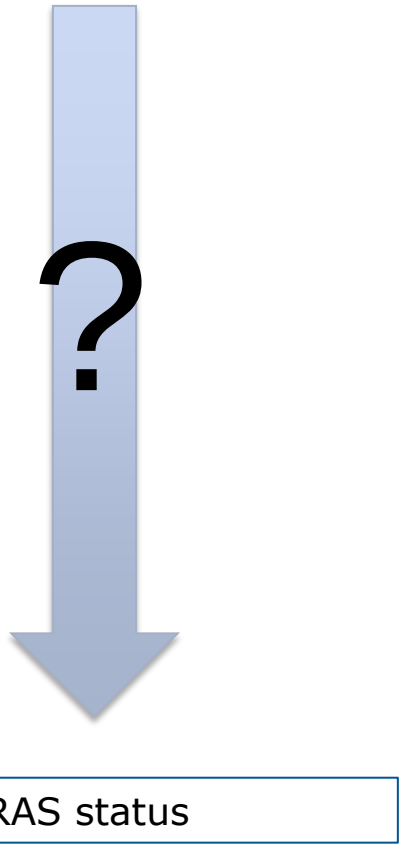
Draw conclusion for each report

- positive KRAS test
- negative KRAS test
- KRAS test requested/performed, but no result in report
- no relevant information

Case study 1 – KRAS in colorectal cancer

Methodology

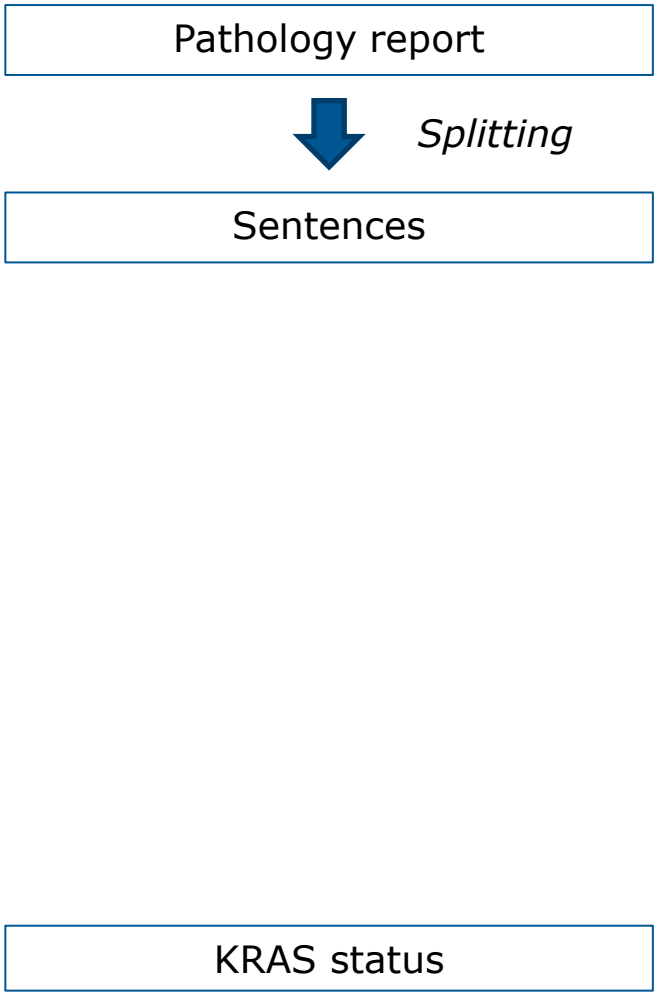
Chain of text mining tools



Case study 1 – KRAS in colorectal cancer

Methodology

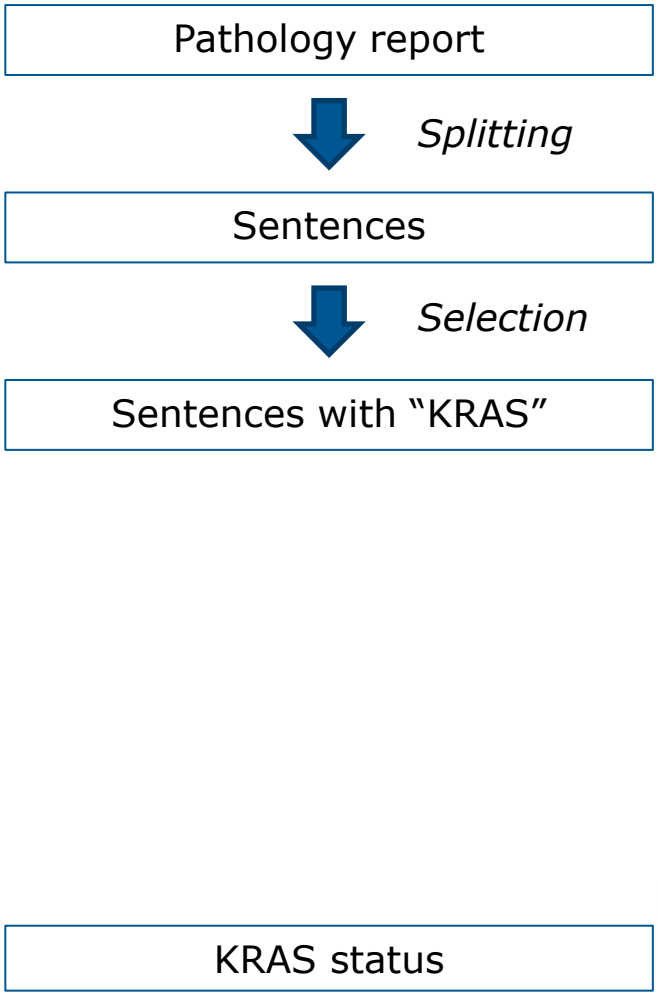
Chain of text mining tools



Case study 1 – KRAS in colorectal cancer

Methodology

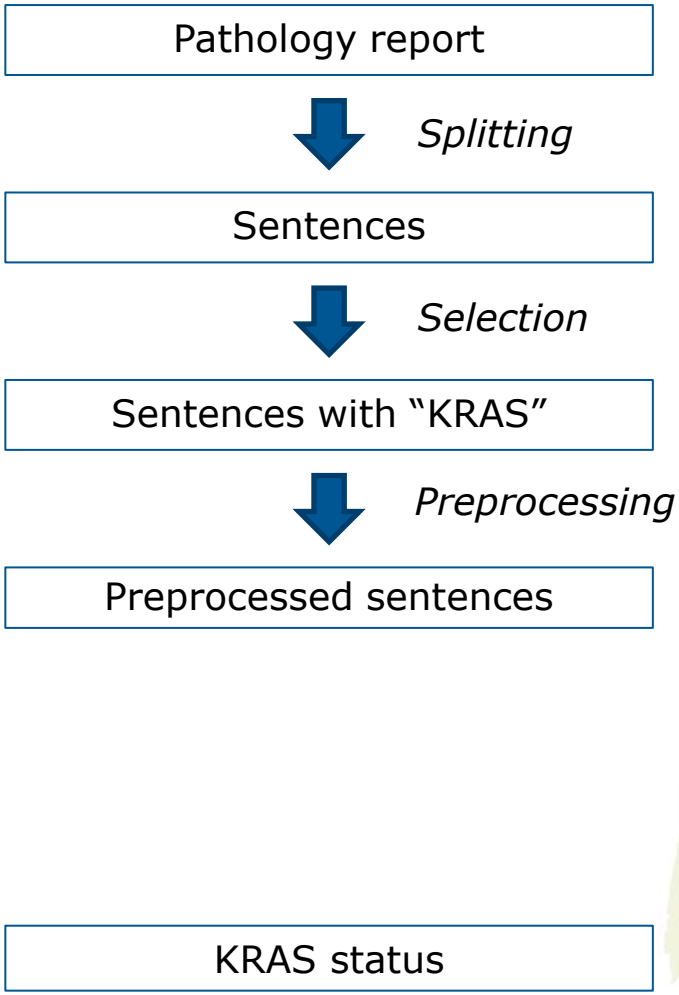
Chain of text mining tools



Case study 1 – KRAS in colorectal cancer

Methodology

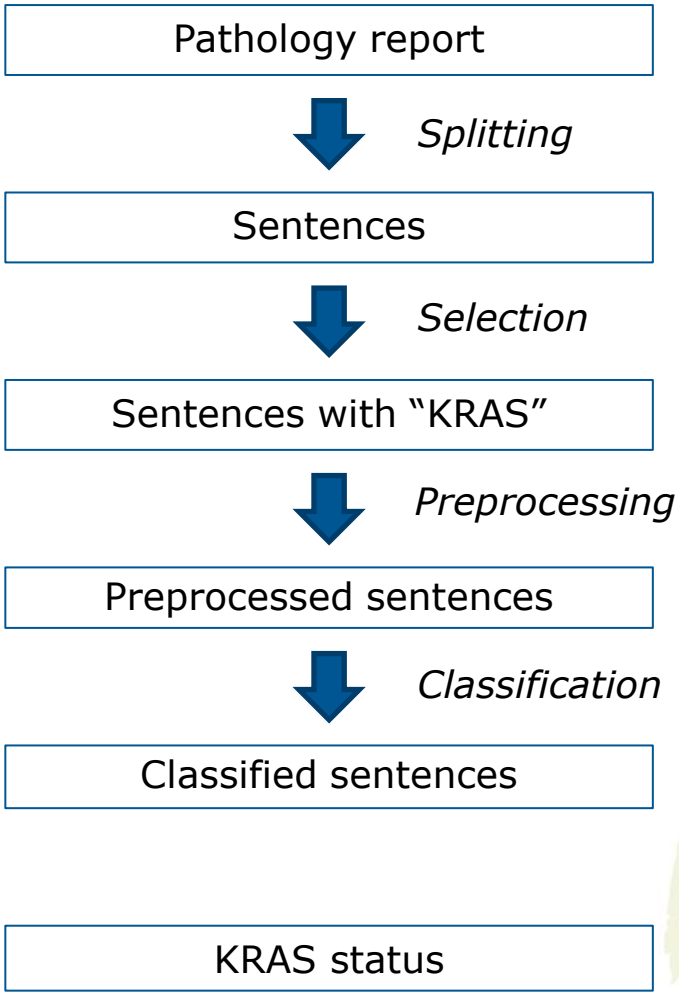
Chain of text mining tools



Case study 1 – KRAS in colorectal cancer

Methodology

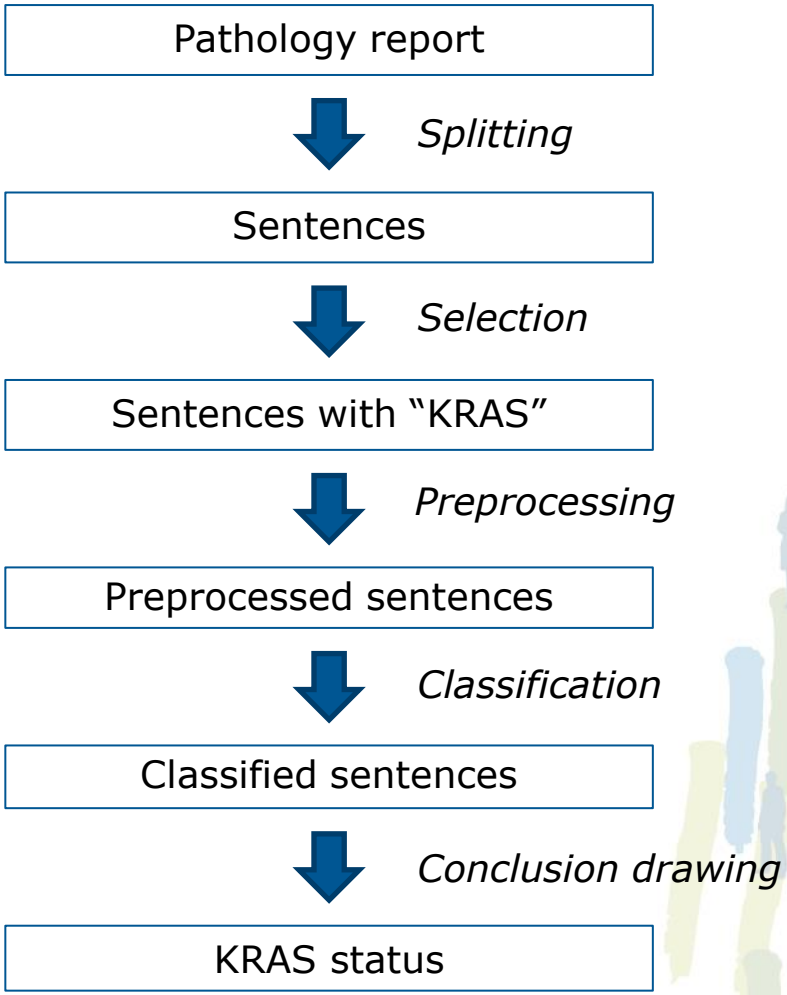
Chain of text mining tools



Case study 1 – KRAS in colorectal cancer

Methodology

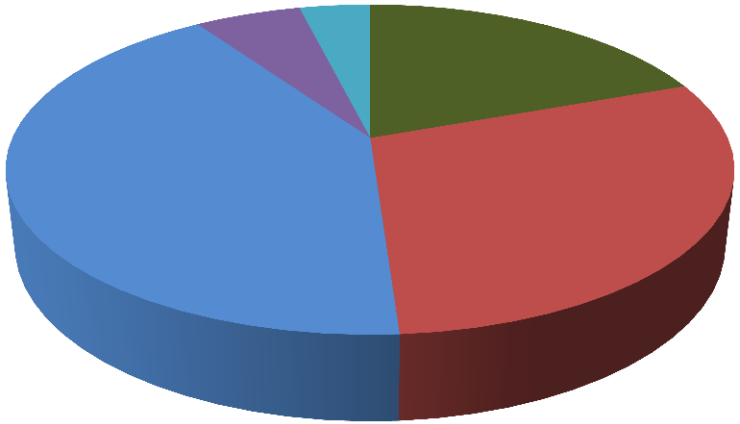
Chain of text mining tools



Case study 1 – KRAS in colorectal cancer

Results

11,446 colorectal cancer reports



46% clear KRAS test result

32% KRAS mutation

68% no KRAS mutation

Case study 1 – KRAS in colorectal cancer

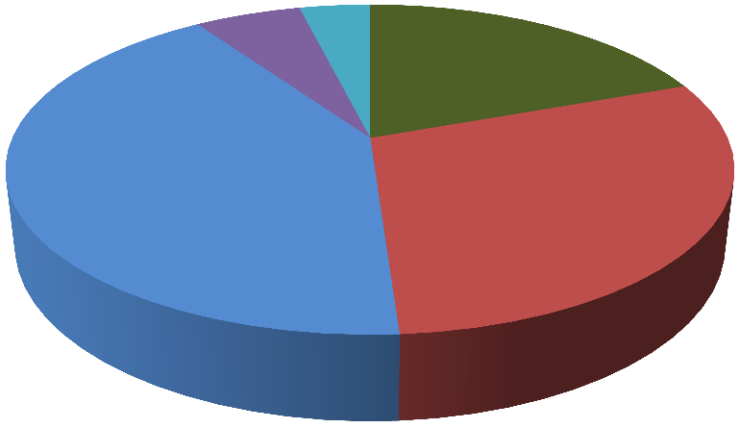
Results

11,446 colorectal cancer reports

46% clear KRAS test result

32% KRAS mutation

44% KRAS test requested/performed, but no result in text



68% no KRAS mutation

Case study 1 – KRAS in colorectal cancer

Results

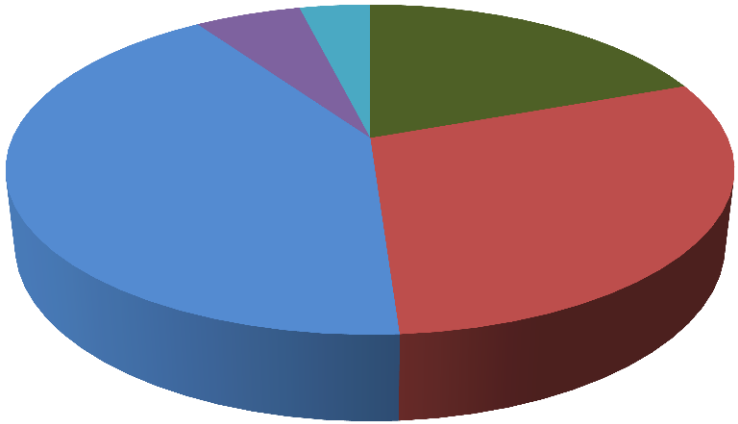
11,446 colorectal cancer reports

4 % contradictory results

6% no relevant information

46% clear KRAS test result

32% KRAS mutation



44% KRAS test requested/performed, but no result in text

68% no KRAS mutation

Case study 2 - HPV in cervical smears

Introduction

Screening test report

- More than 125,000/year
- Some contain result of HPV test

Text collection

- 163 pathology reports for cervical smears
- 2015-2016

Draw a conclusion for each report

- Result of HPV genetic test is present
- No result of HPV genetic test

Case study 2 - HPV in cervical smears

Methodology

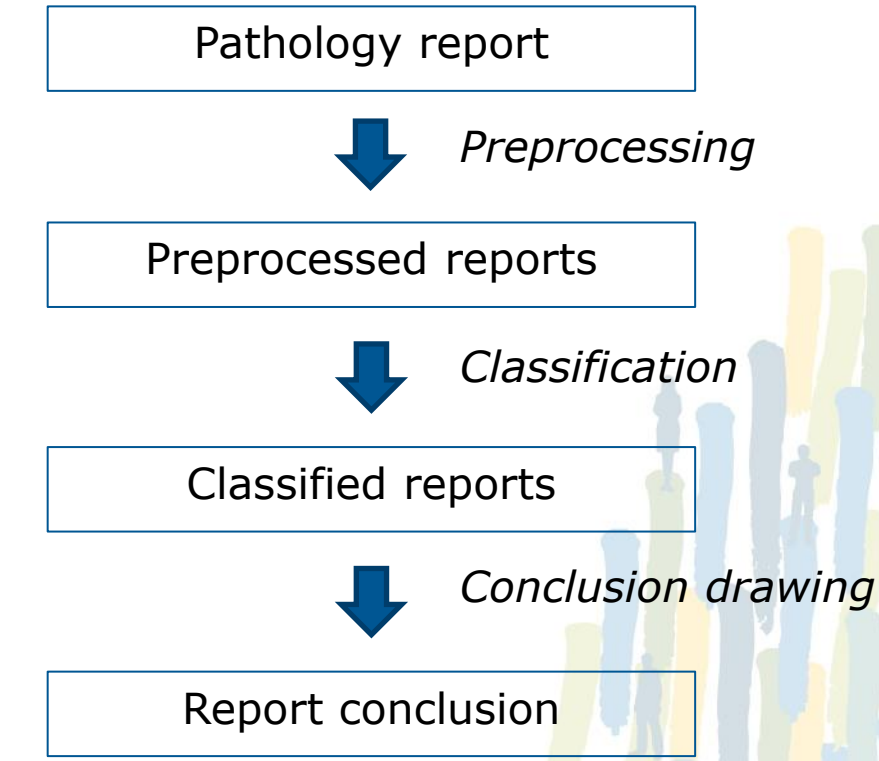
Similar workflow to case study 1

Focus on classification

Train set: 72 reports

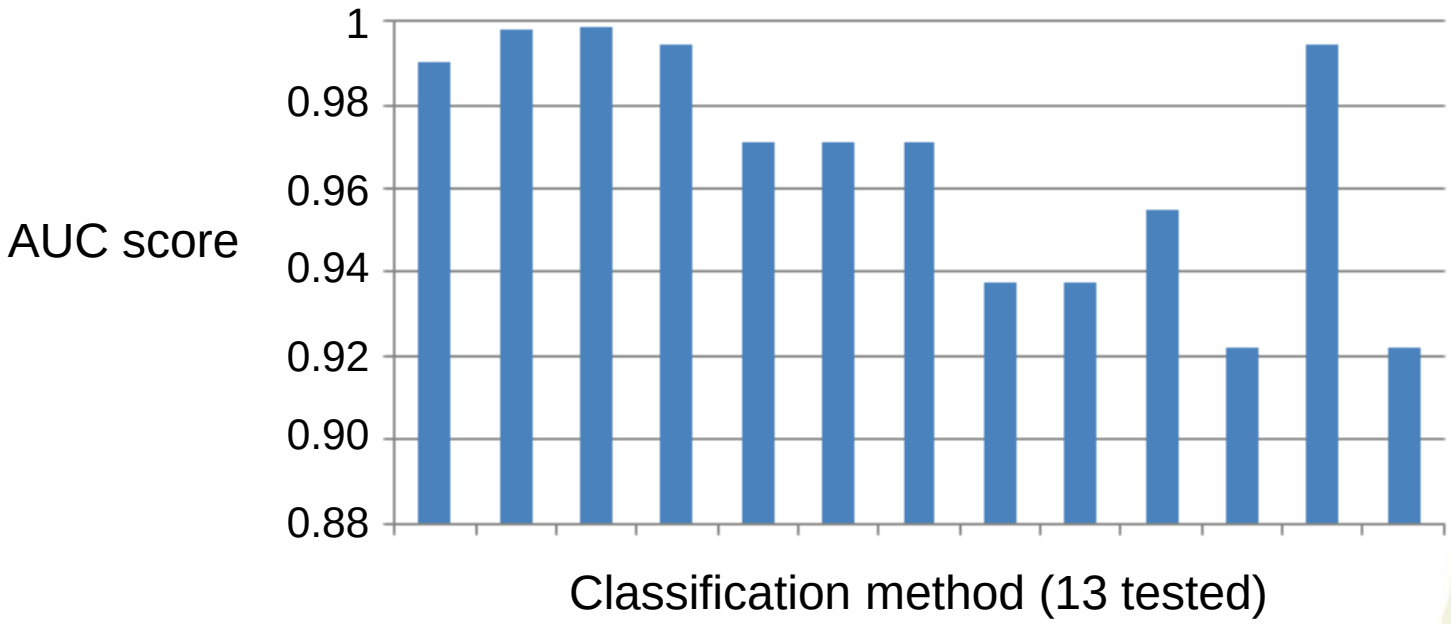
Test set: 91 reports

13 standard machine learning methods



Case study 2 - HPV in cervical smears

Results



Conclusions

Semi-automatic information extraction from pathology reports

- Possible with acceptable accuracy
- Different lesion types
- Different text mining techniques

Further work

- Refinement of methodologies
  - Accuracy
  - Speed
- Enrichment of BCR data with biomolecular information
- Involvement in research projects
  - Mammaprint

Stichting tegen Kanker

Fondation contre le Cancer

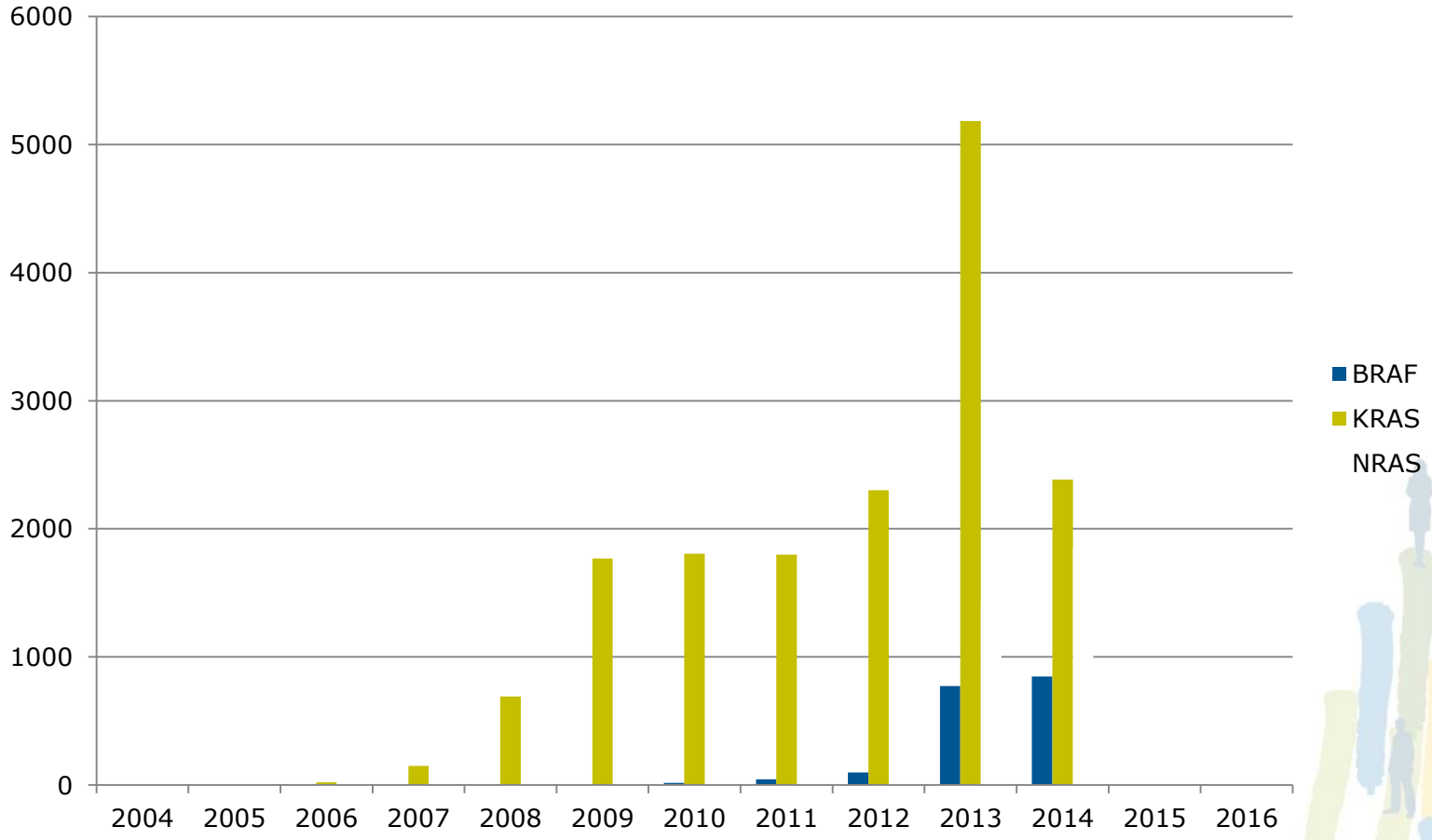
Questions?

Stichting
tegen Kanker

Fondation
contre le Cancer

Back-up slides

Gene count in pathology reports



Case study 1

Pathology report

Splitting

Sentences

Preprocessing

Preprocessed sentences

Classification

Classified sentences

Conclusion drawing

KRAS status

Classification

Manual classification

- 500 most frequent sentences

Automatic classification

- Count words in each sentence -> vector with counts
- Create term-document matrix

	kras	exon	mutation	détecté	…
sentence 214	0	2	1	0	…
sentence 215	1	2	0	1	…
…	…	…	…	…	…

- For each unclassified sentence, calculate distance with all classified sentences
- Take votes from $k = 3$ nearest neighbours

Result

- 14,971 classified, preprocessed sentences
- Accuracy: 76.6%

Case study 1 – KRAS in colorectal cancer

Accuracy

500 manually classifed sentences

train set: 400

test set: 100

77 % correctly classified

Case study 1

Pathology report

↓ *Splitting*

Sentences

↓ *Preprocessing*

Preprocessed sentences

↓ *Classification*

Classified sentences

↓ *Conclusion drawing*

KRAS status

Conclusion drawing

For each report

- Regroup classified sentences
- Draw conclusion

Case study 1

Pathology report

↓ *Splitting*

Sentences

↓ *Preprocessing*

Preprocessed sentences

↓ *Classification*

Classified sentences

↓ *Conclusion drawing*

KRAS status

Conclusion drawing

For each report

- Regroup classified sentences
- Draw conclusion

E.g.

- (O) « Il est a noter qu un resultat négatif n'exclut pas formellement la présence d'une mutation au sein du gène kras. »
- (A) « Compte rendu complémentaire: diagnostic moléculaire: les mutations des codons 12 et 13 du gène kras ont été recherchées par biologie moléculaire. »
- (P) « Conclusion: l'analyse de biologie moléculaire met en évidence une mutation dans le codon 12 du gène kras. »

Case study 1

Pathology report

↓ *Splitting*

Sentences

↓ *Preprocessing*

Preprocessed sentences

↓ *Classification*

Classified sentences

↓ *Conclusion drawing*

KRAS status

Conclusion drawing

For each report

- Regroup classified sentences
- Draw conclusion

E.g.

- (O) « Il est a noter qu un resultat négatif n'exclut pas formellement la présence d'une mutation au sein du gène kras. »
- (A) « Compte rendu complémentaire: diagnostic moléculaire: les mutations des codons 12 et 13 du gène kras ont été recherchées par biologie moléculaire. »
- (P) « Conclusion: l'analyse de biologie moléculaire met en évidence une mutation dans le codon 12 du gène kras. »

-> P

Area under the ROC curve

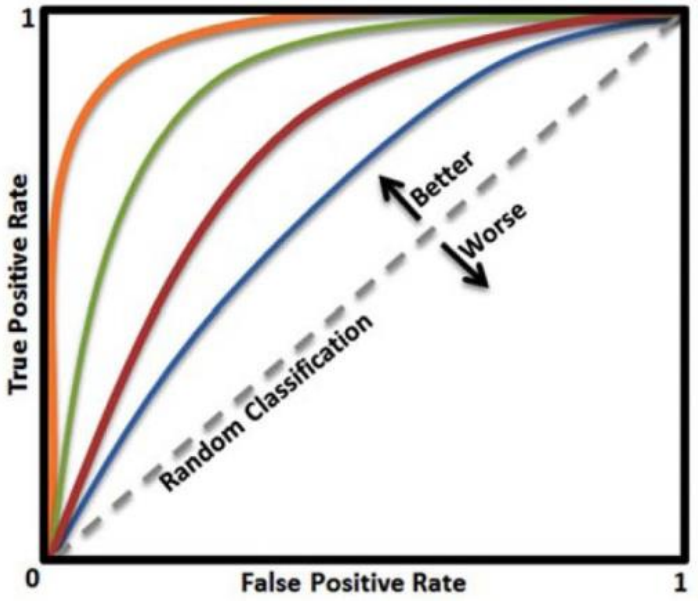
		Predicted	
		+	-
Actual	+	True positive (TP)	False negative (FN)
	-	False positive (FP)	True negative (TN)

True positive rate (recall) TP/ (TP + FN)
False positive rate (fall-out) FP/ (FP + FN)

Combined: ROC-curve (receiver operating characteristic curve)

Area under ROC-curve

1	perfect classification
0.5	random
< 0.5	worse than random



Case study 2 - HPV cervical smears

Results