Contents

Abstract iii

Resumen vii

Resum xi

Glossary xv

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Research questions and objectives . . . 2
- 1.3 Thesis contributions . . . 4
  - 1.3.1 Main contributions . . . 4
  - 1.3.2 Scientific publications . . . 7
  - 1.3.3 Software . . . 9
  - 1.3.4 Other contributions . . . 9
- 1.4 Projects and partners . . . 11
- 1.5 Thesis outline . . . 14

2 Rationale 17
- 2.1 Biomedical data quality . . . 17
  - 2.1.1 Data quality dimensions . . . 18
  - 2.1.2 Multi-source and temporal variability . . . 22
- 2.2 Theoretical background . . . 24
  - 2.2.1 Variables and probability distributions . . . 24
  - 2.2.2 Comparing distributions . . . 36
  - 2.2.3 Information geometry . . . 40
  - 2.2.4 Multi-dimensional scaling . . . 44

3 Comparative study of probability distribution distances 49
- 3.1 Introduction . . . 49
- 3.2 Background . . . 50
- 3.3 Methods . . . 51
  - 3.3.1 Simulation . . . 51
  - 3.3.2 Estimation of probability densities . . . 53

3.3.3 Studied distances . . . 54
3.4 Results . . . 55
3.5 Discussion . . . 57
3.6 Conclusions . . . 59

4 Multi-source variability metrics for biomedical data 61
4.1 Introduction . . . 62
4.2 Background . . . 63
4.2.1 Variability in biomedical data . . . 63
4.2.2 Data source variability in the context of Data Quality 65
4.2.3 Dissimilarities between biomedical data distributions 66
4.3 Simplices and properties . . . 66
4.4 Methods . . . 67
4.4.1 Estimation of PDF densities . . . 68
4.4.2 Calculus of pairwise PDF distances . . . 69
4.4.3 Euclidean embedding using multidimensional scaling 69
4.4.4 PDF simplex building . . . 70
4.4.5 Calculus of metrics . . . 71
4.4.6 Multi-source variability (MSV) plot . . . 73
4.5 Evaluation . . . 73
4.5.1 Evaluation of scalability . . . 74
4.5.2 Evaluation on real data (UCI Heart Disease) . . . 76
4.6 Discussion . . . 85
4.6.1 Significance . . . 85
4.6.2 Limitations . . . 86
4.6.3 Future work . . . 87
4.7 Conclusions . . . 89

5 Probabilistic change detection and visualization methods 91
5.1 Introduction . . . 92
5.2 Background . . . 94
5.2.1 Probabilistic distances on biomedical data distributions 94
5.2.2 Change detection . . . 95
5.3 Proposed methods . . . 97
5.3.1 Probabilistic framework . . . 97
5.3.2 Change monitoring . . . 100
5.3.3 Characterization and temporal subgroup discovery . . 103
5.4 Data . . . 104
5.5 Evaluation . . . 107
5.5.1 Change monitoring . . . 107
5.5.2 Characterization and temporal subgroup discovery . . 108
5.6 Discussion . . . 110
5.6.1 Significance . . . 110
5.6.2 Comparison with related work . . . 111

5.6.3 Limitations . . . 113
5.6.4 Future work . . . 114
5.7 Conclusion . . . 114

6 Applications to case studies **119**
6.1 Introductory notes . . . 119
6.1.1 Summary of the applied methods . . . 119
6.1.2 Additional method combining multi-source and temporal variability . . . 122
6.2 Mortality Registry of the Region of Valencia . . . 123
6.2.1 Materials . . . 123
6.2.2 Results . . . 124
6.2.3 Discussion . . . 132
6.3 Other case studies . . . 135
6.3.1 Cancer Registry of the Region of Valencia . . . 135
6.3.2 Breast Cancer multi-source dataset . . . 139
6.3.3 In-vitro Fertilization dataset . . . 140
6.4 Limitations . . . 141
6.5 Conclusions . . . 142

7 Biomedical data quality framework **147**
7.1 Multi-source and temporal variability . . . 148
7.1.1 Systematic approach . . . 148
7.1.2 Developed software toolbox . . . 150
7.2 Towards a general data quality framework . . . 155
7.2.1 Functionalities and outcomes . . . 156
7.2.2 Data . . . 156
7.2.3 Data quality dimensions . . . 159
7.2.4 Axes . . . 162
7.2.5 Measurements of (dimension,axis) pairs . . . 163
7.2.6 Discussion . . . 164
7.3 Derived applications . . . 168
7.3.1 Data quality assured perinatal repository . . . 168
7.3.2 Contextualization of data for their reuse in CDSSs reuse using an HL7-CDA wrapper . . . 170
7.3.3 Qualize . . . 174

8 Concluding remarks and recommendations **177**
8.1 Concluding remarks . . . 177
8.2 Recommendations . . . 181

Bibliography **187**

A Fisher Information Matrix **201**

B Develoment of equations of simplex properties **203**
B.1 Development of Equation 4.2: $d_{1R}(D)$ 203
B.2 Development of Equation 4.3: $d_{max}(D)$ 203

C Supplemental material for Chapter 5 **205**

D Basic examples of the variability methods **209**
D.1 Multi-source variability 209
D.2 Temporal variability 211

E Supplemental material for the Mortality case study **217**
E.1 WHO ICD-10 Mortality Condensed List 1 217
E.2 Sample size tables 219
E.3 Temporal heat maps of intermediate cause 1 and 2 223
E.4 Unfilled values by Health Department 224
E.5 Multi-site variability of age of death 225
E.6 Dendrograms of initial cause 1 and intermediate cause 2 . . 226
E.7 Spanish Certificates of Death in the period 2000-2012 227
E.8 Temporal variability of basic cause of death 230
E.9 Temporal heatmaps of age at death 230