Subgroups of Gastric Cancer Patients Characterized with an Integrated Large Biomarker Datasets using Association Rules

Williams C1, Adamczyk B2, Afshar M1, Kamali-Moghaddam M3, Karlsson NG2, Guergova-Kuras M1, Lisacek F1, Mereiter SS6, Morley D1, Parmentier F1, Polom K1, Rovelli F7, Reis CA5,6,9,10, Shen Q3 and Tognetti Y1

1. Ariens Pharmacuticals, Paris, France; 2. Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Sweden; 3. Department of Immunogenetics, Genetics and Pathology, Science for Life laboratory, Uppsala University, Sweden; 4. Proteomics Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland; 5. Institute of Investigation in Haematology, University of Porto, Portugal; 6. KITPDM - Institute of Molecular Pathology and Immunology, University of Porto, Portugal; 7. Department General Surgery and Surgical Oncology, University of Porto, Portugal; 8. University of Medicine, University of Porto, Portugal; 10. Instituto de Ciências Biomédicas Abel Salazar, University of Porto, Portugal

Abstract

Gastric cancer (GC) is one of the deadliest forms of cancer worldwide, partly due to the lack of early diagnosis. Availability of molecular data, characterizing cancer patients and their tumour, is required for improved diagnosis and prognosis of patients. The committee of clinicians to provide a precise medicine approach in the diagnosis, prognosis and treatment of GC drives the need for better biological markers. We describe a retrospective study collected proteomic, immunohistochemistry, Helicobacter pylori, and blood biomarker measurements from tissue and serum samples of 307 gastric cancer patients that underwent surgery in the Division of Surgical Oncology, at Tertiary University Hospital of Siena, Italy. In this work, we developed a specific framework dedicated to the integration of multiple datasets from several heterogeneous sources and platforms. Experimental data was integrated with clinical, historical and survival information available for patients providing a large heterogeneous database of 848 variables. This study identified subgroups of patients of clinical importance using a Machine Learning methodology (KEM, Knowledge Extraction and Management®) that provides, through exhaustive exploration of all relationships between patient’s variables, an hypothesis-driven approach helping interpret this broad database and thus identify actionable hypotheses. We systematically extracted all logical associations between experimental measures and clinical outcomes obtaining a knowledge base of over 1000 associations identifying potential disease risk markers.

3. Results

LYMPHNODE METASTASIS

M1, M2, Molecular Subtypes

VALIDATION: Continuing previously described relationships

1. Metric Filtering

The rule of interest found by filtering by metrics (circled here on left-hand side) was found in a predictive direction (Var − Outcome) in the semantic filtering subset: M1status (High) → Nodes (ND)

2. Semantic Filtering

• Most interesting: Invasive Growth, High Node, Perineural invasion

SIMPRESSING STAGING

DISCOVERY: Uncovering new relationships

1. Metric Filtering

Support ≥ 4, Confidence ≥ 65%, Lift ≥ 2.2

2. Semantic Filtering

In a second step the role of interest, circled above, was checked if it was present in a predictive direction (Var − Outcome) in the semantic filtered rules subset: DCBLLD2.S.a (High) → Simp. Stage (1)

3. Conclusions

KEM® platform helps generate new hypotheses and validate previous knowledge from associations. This work describes a data-driven framework using association rules to extract knowledge from an integrated database. A subset of identified relationships were presented and discussed. This work demonstrates the potential of combining powerful Machine Learning tools, experimental glycomic, proteomic data and clinical information to discover potential markers for non-invasive diagnosis and prognosis of gastric cancer.