Call identifier: H2020-ICT-2016 - **Grant agreement no**: **732907**

Topic: ICT-18-2016 - Big data PPP: privacy-preserving big data technologies

Deliverable 8.5

Value Estimation Model

Start of the project: 1st November 2016

Ending Date: or 31st of December 2019

Partner responsible for this deliverable: **HES-SO**

Version: **1.1**

Document Classification

Title	**Value Estimation Model for Datasets**
Deliverable	D8.5
Reporting Period	01.05.2018 - 31.10.2019
Authors	Matthew Jeffryes, Douglas Teodoro, Patrick Ruch
Work Package	WP8
Security	Public
Nature	Demonstrator
Keyword(s)	Relative value estimation; Mutual information; k-Nearest Neighbors

Document History

Name	**Remark**	**Version**	**Date**
Matthew Jeffryes	First Version	0.1	11.08.2019
Matthew Jeffryes	Submitted for review	1.0	12.11.2019
Matthew Jeffryes	Revised after review	1.1	04.12.2019

List of Contributors

Name	**Affiliation**
Douglas Teodoro	HES-SO
Patrick Ruch	HES-SO
Romain Tanzer	HES-SO
Valentine Rech de Laval	HES-SO

List of reviewers

Name	Affiliation
Andre Aichert	SHS
Davide Zaccagnini	Lynkeus
Antonella Trezzani	Lynkeus

Table of contents

1 Preface

This document describes the work performed in the context of WP8 “Prototypical Implementation of Advanced Data Analytics (big data) Use Case on Securely Anonymised and Encrypted Data”. The report provides a description of the way the relative value of a particular subset of data can be estimated when compared to the value of the full dataset it belongs to. The work performed is mostly related to the task T8.4, which started in M18 and finished in M38.

2 Executive Summary

2.1 Scope

The MyHealthMyData (MHMD) platform recognizes four main stakeholders in the data value chain - hospitals, citizens, research centres and industry - with different interests. While citizens and hospitals will share very heterogeneous and privacy sensitive datasets in the network, research centres and industry need streamlined and homogeneous ways to search, discovery and access these datasets.

In this context, WP8 provides services to tentatively estimate whether a particular subset of a dataset is relatively more valuable from an information theoretic point of view than other subsets of the same datasets. The main application scenario consists of computing how a non-shared subset of data could potentially be generated from the shared subset from an information theoretic perspective.

2.2 Problem being addressed

When citizens share sensitive data with other stakeholders, they may unknowingly disclose more of their personal information than they intend to.

2.3 Scientific approach and work undertaken

Using a dataset from PubMed as a stand in for patient data, we have developed a system which, given a set of data to be shared, identifies the unshared data which is most strongly associated with it. The performance of this system has been quantified by its recall and mean average precision.

2.4 Achievements

The system can be accessed via a simple web user interface or a JSON API. The system, as developed, is able to recall unshared data. If applied to patient data this could be used to inform patients of the hidden information which they are unknowingly sharing.

2.5 Relationship to the rest of the project

In the context of the MHMD project, the system developed in this work package can be used to keep patients informed about the extent of the data which they are sharing. By proactively warning them about possible inadvertent disclosures, they will feel more comfortable entrusting other stakeholders with their sensitive information. This reassurance will help to transform “disengaged, sceptical users” into “active and motivated patients with full control over their data” as envisioned in the MHMD project ambitions.

2.6 Conformance to the “Description of work”

The description of work requires that the system accessible to end users such as patients and data providers via a web interface or a JSON API, will output a list of “complementary data” with a “confidence estimate scale”. The implemented system can be accessed by web interface or JSON API. It outputs complementary data, scored by mutual information value.

2.7 Next steps

The system should be tested with data in a closer format to the patient data which will be applicable to MHMD.

3. Introduction

Data shared can carry more information than might be initially evident, especially for layman users willing to share content from their own Electronic Health Records (EHR). Similarly, healthcare institutions (Research or teaching hospitals, General Practitioners, ...) willing to engage into a cooperation with data analytics partners (Public or Corporate Research) may not be in a position to adequately measure the relative importance of a given subset of the data they own.

According to the GDPR, informed consent is only possible when citizens are able to understand the consequences of sharing their personal data. However, it is often difficult to turn such an understanding into actionable knowledge.

Many patients may have little concern about sharing their medical diagnoses. But in some instances, a patient may want to keep details of their medical history confidential. For example, an HIV+ patient may want to keep information about their diagnosis more closely held. In the My Health My Data Data Catalogue, such a patient would be able to control access to their diagnosis. However, they may not realise that other information beyond the fact of the diagnosis could allow someone else to infer their status. Most HIV patients are treated with anti-retroviral therapy. Therefore, if the patient shares their prescription history, they may inadvertently disclose their HIV status. Similarly, HIV patients may be treated by drugs targeting HIV comorbidities (e.g. Kaposi's sarcoma), which may thus indirectly disclose the HIV positivity. The general idea is thus to inform the patient about how associated two subsets of the data they are ready to share are. A data provider might then think twice before sharing a certain set of drug prescription once he is informed that by sharing such a dataset, about 60% of the diagnosis he wants to keep hidden can be automatically recovered.

4. Clinical cases

In order to demonstrate the possibility of hidden information being disclosed, we are using a data set which we believe has similar characteristics to patient data, but which avoids the complication of using real human subject data, which entails strict controls on handling. We thus, used a set of published clinical cases as found in the MEDLINE digital library.

As an analogue to medical encodings (diagnosis, drug prescription, surgery or diagnosis procedures, ...) associated with EHR, we are using the Medical Subject Headings (MeSH) assigned to biomedical literature by the US National Library of Medicine. The present solution is inspired by (Ruch et al. 2008), which used different sections (anamnesis, diagnosis, prescription, ...) of clinical narratives to automatically assign ICD-10 diagnosis codes for sake of billing. MeSH is a hierarchical vocabulary of terms which are attached to entries in the PubMed citation database, based on the major themes of the publication. For example, an epidemiology study on the influenza vaccine might be assigned the MeSH terms “Human Influenza”, “Health Policy”, “Vaccination”. MeSH headings are part of a tree like structure. For example, “Human Influenza” is part of the disease tree, and is categorised under “Respiratory Tract Infections”.

Just as a particular drug and disease may have a strong association, so too may their MeSH terms. We would expect the MeSH terms for “Human Influenza” and “Oseltamivir” (Tamiflu) to appear together quite frequently, we used the MeSH terms covering diseases, drugs and medical procedures as a testbed for investigating how hidden information can be identified and presented.

5. Sample results

A fundamental way of quantifying the “hidden” information contained within other data is *mutual information*. Formally, this is a measure of the amount of information which one random variable contains about another random variable. If two events are completely independent, for example the results of flipping two different coins, then they have zero mutual information. Mutual information rises as the two variables provide more information about each other.

In the case of the example dataset, we can treat the presence or absence of a particular MeSH term attached to a publication as being a binary random variable. The mutual information (MI) for a pair of MeSH terms will increase if they appear very often together, but infrequently without each other. MI is especially relevant to identify content-bearing co-occurrences (see e.g. Stolz 1965) as they only require large corpora and are mostly language independent.

We calculated the mutual information between pairs of MeSH terms which have appeared together in PubMed in the PubMed baseline data files numbered 0900 to 0973 which roughly corresponds to 2018 to early 2019. In these data there are 12,102 distinct terms describing a drug or chemical, a disease or diagnosis, or a treatment or diagnostic procedure assigned to publications, and 1,600,525 distinct pairs of terms have been assigned to the same publication at least 3 times. Shown in figure 1 is the distribution of the higher mutual information values, for each of the possible combinations of categories of drug, disease or procedure.

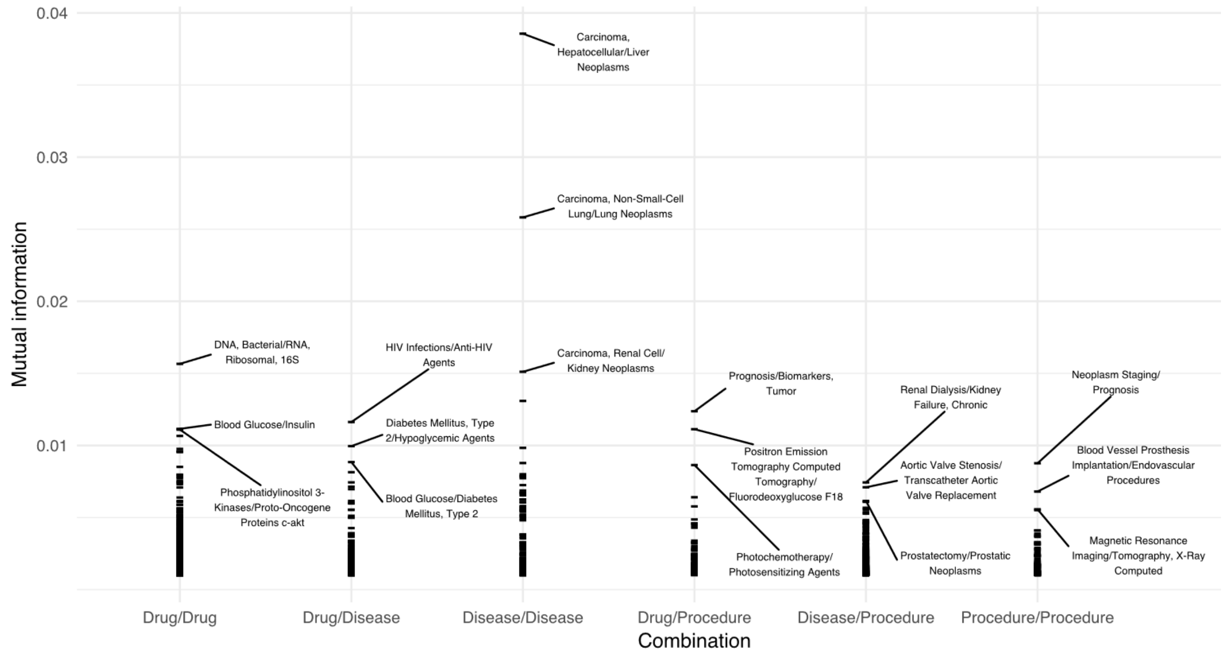

Figure 1: The distribution of the mutual information of every pair of terms in the MeSH data set falling into the drug, disease or procedure categories. The descriptions for the 3 pairs with the highest mutual information in every combination of category is shown.

The pairs with the highest mutual information appear very naturally related. For example, the disease/procedure pair of *Chronic Kidney Failure* and *Renal Dialysis* and the disease/drug pair of *Type 2 Diabetes Mellitus* and *Hypoglycemic Agents*.

6. Web service description

We have developed a web service which will provide access to estimation of the hidden information encoded in a particular set of disclosed data. The web service can be loaded with arbitrary data in a given set of categories. As an example data set, we are again using the MeSH data set. In this case, a user can submit via a web page or REST API the terms which are to be shared, and the service returns the possible hidden information which these terms encode, broken out by category. For example, querying with the MeSH term for Kaposi's sarcoma returns as its highest scoring diseases *Skin Neoplasms* and *HIV Infections*, and its highest scoring procedure *Highly Active Antiretroviral Therapy*. In addition to the terms with the highest mutual information, the service also returns the most similar documents to the query. An example of the JSON format returned by the web service is given in the appendix.

To evaluate the web service, we have sampled from the PubMed baseline files numbered 0890 to 0899, which contain 300,000 citations. From this, we identified citations assigned at least one MeSH term from each of the three categories of drug, disease and procedure. There are 2,185 such citations. From these, we sampled 30 citations and used their MeSH terms to query the web service. For each category, the ability to identify the terms within it using the terms from the other two categories was analysed. That is, for the drug category, the terms in the procedure and disease categories were used to query the web service, and so on. The performance of the service over these queries is shown in Table 1 and Figure 1.

Category	**Top 20 recall**		**Mean average precision**
	Micro	**Macro**	
Drug	0.451	0.439	0.254
Disease	0.611	0.576	0.255
Procedure	0.841	0.811	0.417
All	*0.635*	*0.571*	*0.309*

Table 1: Top 20 recall and mean average precision for 30 queries against the web service.

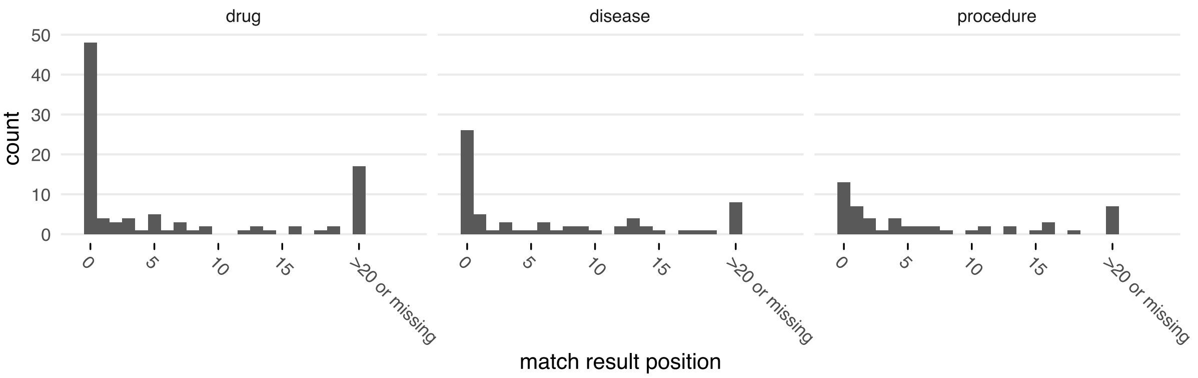

Figure 1: Histogram showing the position within the search results for each query. In cases where there were multiple correct results for a query (i.e., a publication had more than one drug, disease or procedure MeSH term assigned to it), each correct result has contributed towards the count for the position at which it appeared.

7. Discussion

Out of the three axis of classification selected in the study, Table 1 shows that procedures provide the highest relative value, and drug prescription & diagnosis the lowest relative association value with the other entities - diseases and drug. Basically, it means that when a given dataset owner decides to share a subset of data, the prescription contains the highest relative value. If the dataset owner decides to share this subset, the dataset recipient can recover the non-shared diagnosis and drug subsets with more than 80% recall.

By way of example, the publication “Multitarget stool DNA tests increases colorectal cancer screening among previously noncompliant Medicare patients” (PM28210082) is assigned the *disease* MeSH terms “DNA, Neoplasm” (D004273) and “Colorectal Neoplasms” (D015179). The MeSH term with the highest mutual information with any of these terms is “Colonoscopy” (D003113) which is indeed assigned to the publication as a MeSH term in the *procedure* category.

As an example of a failure to identify a term, the publication “Combining doxorubicin with a phenolic extract from flaxseed oil: Evaluation of the effect on two breast cancer cell lines” (PM28101573) is assigned the MeSH term “Carcinoma, Ductal” (D044584), which is not found within the top 20 matches for the publication’s *procedure* or *drug* query terms. However, “Breast Neoplasms” (D001943) is the fifth most related term, and in some contexts may be a close enough “diagnosis” in the context of patient privacy. The effectiveness of this system therefore depends on how the data it is used with, is encoded and the degree to which fuzzy matches are acceptable.

8. Conclusion

We tested different methods to assess the relative value of a given subset, including one approach based on mutual information and one inspired by information retrieval document to document distance. The service and API is available.

9. References

Ruch P, Gobeill J, Tbahriti I, Geissbühler A: From Episodes of Care to Diagnosis Codes: Automatic Text Categorization for Medico-Economic Encoding. AMIA Annu Symp Proc. 2008; 2008: 636–640.

Stolz, W. (1965). A probabilistic procedure for grouping words into phrases. Language and Speech, 8.

10. Appendix

Example query

A query to the web service for *disease* terms associated with the *procedure* “Renal Dialysis” (D006435). The list of similar documents and similar terms has been truncated to four from the 100 results returned by the API.

```
GET /api/similar/disease/D006435

{
  "similar_documents": [
    {
      "pmid": "4485908",
      "score": 8.506218,
      "terms": [
        {
          "name": "Hepatitis B",
          "term": "D006509"
        }
        ]
      },
      {
        "pmid": "62937",
        "score": 8.506218,
        "terms": [
          {
            "name": "Hepatic Encephalopathy",
            "term": "D006501"
          }
        ]
      },
      {
        "pmid": "669461",
        "score": 8.49463,
        "terms": [
          {
            "name": "Arteriosclerosis",
            "term": "D001161"
          }
        ]
      },
      {
        "pmid": "4787241",
        "score": 8.484577,
        "terms": [
          {
            "name": "Acute Kidney Injury",
            "term": "D058186"
          }
        ]
      },
    ...
    ],
    "terms": [
      {
        "name": "Kidney Failure, Chronic",
        "query_term": "D006435",
        "score": 0.007435440499262912,
        "term": "D007676"
      },
      {
        "name": "Renal Insufficiency, Chronic",
        "query_term": "D006435",
        "score": 0.002137672378524322,
        "term": "D051436"
      },
      {
        "name": "Graft Occlusion, Vascular",
        "query_term": "D006435",
        "score": 0.0007697885870043375,
      "term": "D006083"
    },
    {
      "name": "Acute Kidney Injury",
      "query_term": "D006435",
      "score": 0.0006404588691324498,
      "term": "D058186"
    },
   ...
  ]
}
```