Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-531

A Knowledge-Base for a Personalized Infectious Disease Risk Prediction System

Retno VINARTI[a,1] and Lucy HEDERMAN[a]
[a] *School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin, Ireland*

Abstract. We present a knowledge-base to represent collated infectious disease risk (IDR) knowledge. The knowledge is about personal and contextual risk of contracting an infectious disease obtained from declarative sources (e.g. Atlas of Human Infectious Diseases). Automated prediction requires encoding this knowledge in a form that can produce risk probabilities (e.g. Bayesian Network – BN). The knowledge-base presented in this paper feeds an algorithm that can auto-generate the BN. The knowledge from 234 infectious diseases was compiled. From this compilation, we designed an ontology and five rule types for modelling IDR knowledge in general. The evaluation aims to assess whether the knowledge-base structure, and its application to three disease-country contexts, meets the needs of personalized IDR prediction system. From the evaluation results, the knowledge-base conforms to the system's purpose: personalization of infectious disease risk.

Keywords. knowledge-base, rules, ontology, infectious disease, risk.

1. Introduction

We envisage a service which predicts a person's risk of contracting an infectious disease (ID) based on their personal attributes (age, diet) and their location (weather, geographical features). This service will supply risk predictions to advisor applications designed to help users take risk-reducing actions (e.g. wear a mask to avoid influenza during a windy week in autumn).

Knowledge about personal and contextual risk of contracting an ID is largely communicated in declarative form; general, stable knowledge is documented in the Atlas of Human Infectious Diseases (AHID) and similar books [1-3]; more specific and up to date knowledge is conveyed in epidemiology journals. Automated prediction requires encoding this knowledge in a form that can produce risk probabilities, such as in a Bayesian Network (BN). In previous work [4] we showed that we can yield accurate predictions by manually encoding the IDR knowledge in a BN. But this approach is not scalable to all IDs, regions and risk factors, nor maintainable to model new knowledge.

Rather than hardcoding current general knowledge for all IDs as a BN, we seek to facilitate the ongoing encoding by epidemiologists of up to date and region-specific ID risk. The knowledge is manually represented by experts as ID risk rules (e.g. a rule that says smoking can double Tuberculosis (TB) risk) defined over a special purpose ontology of ID risk. The knowledge-base (an ontology and collection of ID risk rules) is

[1] Corresponding Author, e-mail: retnor@tcd.ie. Special thanks to Dr. Fariziyah D. Safitri, and Nurul Kodriati, M.Med.Sc., (Ph.D Cand) for feedback on the knowledge-base.

then automatically converted to a BN, using the BN builder algorithm described in [5]. This paper describes the design and evaluation of the knowledge-base.

2. Knowledge-base Design

This section describes the methodology used to design the knowledge-base. First, the ID literature was summarized to identify risk elements and quantitative forms of risk. From this summary, the 'backbone' of the IDR ontology: *person*, *infectious disease* and *environment* was created. Then, five IDR rule types were designed for representing quantitative forms of how risk factors affect an ID risk, defined over the ontology.

The role of the infectious disease risk (IDR) ontology is to capture and organize general IDR knowledge, for both the experts and the BN builder algorithm. An infectious disease is an illness caused by a specific pathogen that results from transmission from infected person, animal or its reservoir to a susceptible human host [9]. From this definition, three entities are involved: (1) pathogen's availability [10-14], (2) transmission method, and (3) susceptible host [15-17]. The compiled summary[2] consists of 234 unique IDs listed in these declarative sources [1-3]; it was clear that *demography*, *behavior* and *environment* are risk elements. Only a few IDs have a *genetic* risk element.

Three main ontology classes were created to the represent IDR 'backbone' specified in this collation: *Infectious disease* class represents an infectious disease name whose risk is being predicted. The *person* class accommodates the personal risk groups for the disease named in the *infectious disease* class. Risk groups describe susceptibility level of host by defining their *demographic* and *behavioral* risk elements. The e*nvironment* class explains the transmission method, pathogen reservoir and availability of the specified disease. This structure (Fig.1) represents a basic semantic structure for general IDs. By default, the ontology instantiated for each ID will contain this structure.

Knowledge about how these risk factors impact risk of a person contracting an ID is encoded as IDR rules over the IDR ontology. The IDR rules are designed to be easy for domain experts to use, while allowing automatic population to a BN's base reasoning. The ID risk knowledge is divided into: (1) *risk ratios* for each risk factor, (2) *prevalence values* for specified regions, (3) *pathogen activity* information during particular climate or location features (e.g. Aedes Aeqypti mosquitoes live at altitudes below 1000m). The IDR rules allow three quantitative forms of knowledge: risk ratios as numerical values, risk tendency as ordinal values, and risk addition or reduction as percentages. The IDR rule types to represent these forms are shown in Table 1, with some declarative examples.

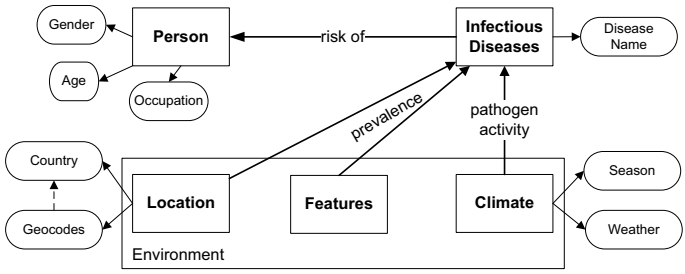

Figure 1. The IDR ontology basic structure with some samples of sub-classes (ellipses)

[2] The summary is available at http://is.gd/IDcompilation

Table 1. The IDR rule types that encode quantitative knowledge forms over the IDR ontology, with examples.

Rule Types	Ontology *class* to encode	Value in data forms	Examples of declarative knowledge to encode
Real Direct Risk Ratios	{risk factors} in *person* or *environment*	Risk ratio in $0 < \mathbb{R} < \infty$	Males have 2.37 times more TB risk than females [19]
`Rule: Person(?all) ^ hasGender(?all, Male) -> alterRisk(TB, 2.37)`			
Real Indirect Risk Ratios	{risk factors} in *person* or *environment*	Risk ratio in %	Fish intake can reduce the TB risk by 50% [20]
`Rule: Person(?all) ^ hasEatingHabits(?all, fish) -> reduceRisk(TB, 50%)`			
Vague Pathogen Status	{location features and climate} in *environment*	Pathogen Activity in Inactive, LessActive, MoreActive	*Mycobacterium tuberculosis* is more active during humid condition [20]
`Rule: Environment(?all) ^ during(?all, humid) -> setPathogen(TB, MoreActive)`			
Vague Risk Ratios	{risk factors} in *person* or *environment*	Risk ratio in High, Low, Medium, n-fold	People who have low body mass index are at a high risk [20]
`Rule: Person(?all) ^ hasBMI(?all, low) -> estimateRisk(TB, high)`			
Real Prevalence	{location and specific features} in *environment*	Prevalence rate in %K or %	TB prevalence in Africa is 395 per 100,000 population.
`Rule: Environment(?all) ^ hasCountryName(?all, Indonesia) -> setRisk(TB, 0.395)`			

3. Evaluation

Evaluation of the knowledge-base aims to assess whether the ontology and rules meet the requirement of the personalization system. Current approaches are evolution-based, logical-based, and metric-based [21]. This evaluation was based on the ontology and rule types described above being instantiated for three disease-country contexts: TB-Africa (34 rules), Dengue-Indonesia (23 rules) and Cholera-India (18 rules) representing air-borne, vector-borne and food/water-borne ID, respectively [18, 26, 27].

An e*volution-based* approach evaluates the ontology based on the changes that may happen. A good ontology is able to accommodate *changes* without reconstructing the basic ontology structure [21, 22]. In the IDR ontology, *changes in domain* happen when new risk factors are found which the knowledge engineer needs to represent in the ontology. *Changes in conceptualization* happen as result of different perspectives between experts. Informally, a GP and an epidemiologist were asked to advise on the IDR ontology: the changes were specific to risk details. Both kinds of changes occurred, but the ontology basic structure remains the same. *Changes in the explicit specification* occur when an ontology is translated from other knowledge representation forms. Since the IDR ontology was built from ID literature, these changes do not occur.

The *logical-based* approach evaluates IDR rules based on rule anomalies that usually occur in rule-bases. We used anomalies defined by COVER [23, 24]: *unused inputs, unsatisfiable condition, unusable consequent, duplicate, circular* and *contradictory rules.* All subclasses are created for the purpose of describing ID risk; therefore, the *unused inputs* anomaly cannot happen. All antecedents and consequents of IDR rules refer to different classes, thus, there are no *circular rules*. For the three disease-country contexts implemented, there are no *unsatisfiable condition* and *unusable consequent* anomalies. However, these may happen when the knowledge-base management system has no integrity checking (e.g. renaming the instances after defining rules). *Contradictory rules* happen when using more than one source to describe the IDR knowledge (e.g. one source says, males have higher TB risk than females, another says males have lower risk). In this case, the epidemiologist is asked to specify a priority [0-

1]; the rules with the highest priority are used for BN building. *Duplicate rules* happen when the same risk ratio is expressed using different rule types (e.g. increase risk by 285% using *Real Indirect Risk Ratios*, and risk ratio 2.85 using *Real Direct Risk Ratios* type). A user interface with grouping feature will be designed to inform experts about similar pre-defined rules at the same disease-country context in the IDR knowledge-base; this should eliminate entry of duplicate rules.

The *metric-based* approach evaluates the ontology using OntoQA metrics [25]: *class richness, class importance* and *relationship richness*. The metrics evaluate the placement of instances within the ontology and the knowledge-base effectiveness. The *class richness* shows the percentage of unused sub-classes; the lower the percentage, the more effective the class. The *class importance* infers the importance based on the dispersion of number of instances. The *relationship richness* shows: (1) for the *infectious disease* class, how many rule types are utilized; (2) for the *person* and *environment* class, how many relations are used. Looking at *class richness* in Table 2, TB is affected by personal (no subclasses unused), rather than environmental risk factors (50% unused). The *class importance* confirms this finding as *person* class for TB is the highest (81.8%); and *environment* class for Cholera has the highest value (54.84%). With regard to *relationship richness*, the *person* and *infectious disease* class have higher percentage (66.67% and 80%) than *environment* class. This shows that the IDR is capable of personalized decisions; and four of five rule types are used to express IDR knowledge.

Table 2. Results of each OntoQA metric for each main class of the IDR[3]

Metrics	Class	Results (in %)		
		Tuberculosis	Dengue	Cholera
Class Richness	Person	0/7 = 0	0/9 = 0	0/4 = 0
	Environment	2/4 = 50	0/8 = 0	0/7 = 0
	Infectious Disease	0/1 = 0	0/1 = 0	0/1 = 0
Class Importance	Person	27/33 = **81.8**	24/46 = 52.17	13/31 = 41.93
	Environment	5/33 = 12.15	21/46 = 45.65	17/31 = **54.84**
	Infectious Disease	1/33 = 3.03	1/46 = 2.17	1/31 = 3.22
Relationship Richness	Person	6/9 = **66.67**	9/17 = 52.9	5/11 = 45.45
	Environment	3/9 = 33.33	8/17 = 47.06	6/11 = 54.54
	Infectious Disease	2/5 = 40	4/5 = **80**	4/5 = **80**

4. Conclusion and Further Works

This paper has presented a knowledge-base for encoding infectious disease risk knowledge which is used in a personalized IDR prediction system. The basic structure of the knowledge-base consists of an ontology and five rule types that represent IDR knowledge for all IDs. Three approaches to knowledge-base evaluation have been applied. Changes are unavoidable in the ontology evolution; however, none of the changes have impact on the ontology basic structure. Four out of six anomaly types are possible in the IDR knowledge-base, however, only one of them is caused by overuse of IDR rule types. Based on three tested cases, the metric-based approach shows that (1) most classes are effective, (2) the ontology is centralized at *person* and *environment* classes; both are equally important for modelling IDR knowledge, (3) high utilization in *person* and *infectious disease* classes confirm the system's purpose: personalization of ID risk prediction.

[3] Details of the TB-IDR, Dengue-IDR and Cholera-IDR can be found in https://is.gd/IDRforMIE

References

[1] H. F. L. Wertheim, P. Horby, Atlas of Human Infectious Diseases, Oxford: Wiley-Blackwell, 2012.
[2] CDC, "Emerging Infectious Diseases," Centers for Disease Control, 2017. https://wwwnc.cdc.gov/eid.
[3] WHO, "Infectious Diseases," WHO, 2017.http://www.who.int/topics/infectious_diseases/factsheets/en/.
[4] R. A. Vinarti and L. M. Hederman, "Personalization of ID Risk Prediction towards automatic generation of a Bayesian Network," in IEEE Computer-based Medical Systems, Thessaloniki, Greece, 2017.
[5] R. A. Vinarti and L. M. Hederman, "Introduction of a BN Builder Algorithm: Personalized Infectious Disease Risk Prediction," in 11th International Health Informatics, ACM, Funchal, Madeira, 2018.
[6] J. Ralyte, X. Franch and S. Brinkkemper, "Advanced Information Systems Engineering," in 24th International Conference, CAiSE 2012, Gdansk, Poland, 2012.
[7] A. Ruttenberg, A. Goldfain, A. Diehl, "Infectious Disease Ontology," The National Center for Biomedical Ontology. http://purl.obolibrary.org/obo/ido.owl.
[8] A. Third, "BioPortal: CARRE Ontology," 2014. https://bioportal.bioontology.org/ontologies/CARRE.
[9] M. L. Barreto, M. G. Teixeira and E. H. Carmo, "Infectious diseases epidemiology," Journal of Epidemiology in Community Health, vol. 60, pp. 192-195, 2006.
[10] N. C. Stenseth, N. I. Samia, "Plague dynamics are driven by climate variation.," vol. 103, no. 35, 2006.
[11] S. M. Upadhyayula, S. R. Mutheheni and H. K. Nayanoori, "Impact of weather variables on mosquitoes infected with Japanese encephalitis virus in Kurnool district, Andhra Pradesh.," vol. 5, no. 5, 2012.
[12] D. Onozuka and M. Hashizume, "Effect of weather variability on the incidence of mumps in children: a time-series analysis.," vol. 139, no. 11, 2011.
[13] W. F. Petersen, "Tuberculosis Weather and Resistance *," 1942.
[14] C. Lau, P. Weinstein and D. Slaney, "Imported cases of Ross River virus disease in New Zealand - a travel medicine perspective.," vol. 10, no. 3, 2012.
[15] A. Fares, "Seasonality of TB of Global Infectious Diseases, vol. 3, no. 1, pp. 46-55, 2011.
[16] H. W. Hethcote, "The Mathematics of Infectious Diseases*," vol. 42, no. 4, 2000.
[17] T. Ditsuwan, T. Liabsuetrakul and V. Chongsuvivatwong, "Assessing the Spreading Patterns of Dengue Infection and Chikungunya Fever Outbreaks in Thailand Using a GIS," vol. 21, no. 4, 2011.
[18] P. Gustafson, V. F. Gomes, C. S. Vieira, "TB in Bissau: incidence and risk factors in an urban community in sub-Saharan Africa," International Journal of Epidemiology, vol. 33, pp. 163-172, 2004.
[19] D. Guwatudde, M. Nakakeeto, E. C. Jones-Lopez, "TB in Household Contacts of Infectious Cases in Kampala, Uganda," American Journal of Epidemiology, vol. 158, no. 9, pp. 887-898, 2003.
[20] Q. Wang, Y. Liu, Y. Ma and L. Han, "Severe hypovitaminosis D in active TB patients and its predictors," Clinical Nutrition, pp. 1-7, 2017.
[21] S. Tartir, I. B. Arpinar and A. P. Sheth, "Ontological Evaluation and Validation," in Theory and Applications of Ontology: Computer Applications, Dordrecht, Springer, 2010, pp. 115-130.
[22] N. F. Noy and M. A. Musen, "The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping," International Journal of Human-Computer Studies, vol. 59, no. 6, pp. 983-1024, 2003.
[23] A. Preece, "Evaluating verification and validation methods in knowledge engineering," 2001.
[24] A. Preece and R. Shinghal, "Foundation and application of knowledge base verification," International Journal of Intelligent Systems, vol. 9, no. 8, pp. 683-701, 1994.
[25] S. Tartir, I. B. Arpinar, M. Moore and A. P. Sheth, "OntoQA: Metric-based Ontology Quality Analysis," in IEEE ICDM 2005 Workshop on Knowledge Acquisition. Houston, Texas, 2005.
[26] A. Prayitno, A.-F. Taurel, J. Nealon, "Dengue seroprevalence and force of primary infection in urban Indonesian children population," PLoS: Neglected Tropical Diseases, vol. 11, no. 6, pp. 1-16, 2017.
[27] V. S. Kumar, S. Devika, S. George and L. Jeyaseelan, "Spatial mapping of acute diarrheal disease using GIS and estimation of relative risk using empirical Bayes approach," vol. 5, 2017.