Automating Population Health Studies Through Semantics and Statistics

Alexander New$^{[0000-0001-8369-1473]}$, Miao Qi$^{[0000-0002-2917-0965]}$, Shruthi Chari$^{[0000-0003-2946-7870]}$, Sabbir M. Rashid$^{[0000-0002-4162-8334]}$, Oshani Seneviratne$^{[0000-0001-8518-917X]}$, James P. McCusker$^{[0000-0003-1085-6059]}$, John S. Erickson$^{[0000-0003-3078-4566]}$, Deborah L. McGuinness$^{[0000-0001-7037-4567]}$, and Kristin P. Bennett$^{[0000-0002-8782-105X]}$

Rensselaer Polytechnic Institute, Troy NY 12180, USA
{newa2,qim,charis,rashis2,senevo,mccusj2,erickj4,mcguid,bennek}@rpi.edu

Abstract. With the rapid development of the Semantic Web, machines are able to understand the contextual meaning of data, including in the field of automated semantics-driven statistical reasoning. This paper introduces a semantics-driven automated approach for solving population health problems with descriptive statistical models. A fusion of semantic and machine learning techniques enables our semantically-targeted analytics framework to automatically discover informative subpopulations that have subpopulation-specific risk factors significantly associated with health conditions such as hypertension and type II diabetes. Based on our health analysis ontology and knowledge graphs, the semantically-targeted analysis automated architecture allows analysts to rapidly and dynamically conduct studies for different health outcomes, risk factors, cohorts, and analysis methods; it also lets the full analysis pipeline be modularly specified in a reusable domain-specific way through the usage of knowledge graph cartridges, which are application-specific fragments of the underlying knowledge graph. We evaluate the semantically-targeted analysis framework for risk analysis using the National Health and Nutrition Examination Survey and conclude that this framework can be readily extended to solve many different learning and statistical tasks, and to exploit datasets from various domains in the future.

Keywords: Automated Machine Learning · Semantic Representation · Statistical Data and Metadata Publication · Population Health

1 Introduction

Population health strives to improve the health outcomes of subject groups through the analysis of enormous health-related datasets collected from members of these groups [5]. With the great advancements in data analytics and increasing scope of population health datasets, accurate use of these data and statistics will be required to monitor and improve population-wide health situations. To understand the relationships between population health determinants and outcomes, observational studies are performed on large patient databases.

These databases include electronic health records and ongoing population-wide surveys, such as the National Health and Nutrition Examination Survey (NHANES, [3]) studied here. Run by the National Center for Health Statistics, NHANES examines about 5000 subjects a year and serves as a primary data resource for population health studies. These studies, however, often suffer from a limited scope, and many studies may require the same repeated domain-specific data preparation procedures. The objective of a study might be confined to a single health condition, a small number of risk factors, and a manually-chosen subject cohort.

In this work, we present a framework, semantically targeted analytics (STA), for automatically generating population health statistical analyses. In order to overcome limitations on study scope, we develop a semantic representation for knowledge from key domains: survey design and analysis, health, and data analytics. Integration of each of these domains via a consistent standard [1] is necessary for our system to formulate and answer meaningful questions. We represent our knowledge as a knowledge graph (KG[1]), containing terms defined by domain-specific best-practice ontologies.

When subject cohorts are no longer manually chosen, there is no guarantee that a linear statistical model will be sufficient to explain associations found in population health datasets. Thus, we utilize the supervised cadre model (SCM, [14]), a machine learning technique that automatically discovers informative subpopulations in datasets. For subpopulations within these subpopulations, associations between response variables and features are approximately linear. The SCM has already been applied to predictive analytics and precision population health [13]; in Section 4, we integrate the SCM with STA.

In STA, semantics encodes, captures, and isolates the domain knowledge needed to model study definitions, statistical techniques, and data. A key component is the *knowledge graph cartridge* (hereafter cartridges): an application-specific subgraph of an underlying KG. Cartridges, further described in section 3.2, are a way to express special-purpose, application-specific sub-graphs, to augment the graph for analysis. They are implemented as RDF KGs and enable an automated "plug and play" architecture. Further, cartridges are used either as input when analysts choose to load them to perform a novel risk study, or as output when the study finding is automatically written into them. Our cartridges are sub-graphs that contribute to a larger analysis graph. Additionally, our output cartridges define the results in a way that is consistent with the input cartridges and contribute to the modularity of STA.

To model and represent the components of our cartridges, we built a Health Analytics Ontology (HAO[2]). HAO models the domain knowledge, analytics

[1] Here, KG refers to a graph that describes real world entities and their interrelations while and enumerating the possible classes and relations of these entities. [19]

[2] The HAO is hosted at https://github.com/TheRensselaerIDEA/hao-ontology.

knowledge, and other analytics pipeline components necessary for population health analysis.

The main contributions of this paper are a semantic representation of population health analysis workflows and results as knowledge graph cartridges, the integration of this representation with precision machine learning techniques for the discovery subpopulation-specific risk factors, and the demonstration of how the STA framework enables rigorous investigation of population health problems. Via cartridges, our STA framework can analyze, interpret, and report studies performed on a wide variety of chronic health conditions and potential risk factors. In Section 5, we present and examine the discoveries found by applying STA to the task of subpopulation-specific identification of risk factors associated with prediabetes and increased total cholesterol levels. Our framework successfully identifies risk factors that are not picked up by standard population-level risk analysis.

1.1 Related work

Our primary inspiration for the KG cartridge is the Oracle database systems notion of a data cartridge [7]; similarities can also be found in the theory of modular ontology design [17] and cheminformatics chemical cartridges [12]. Just as in the data cartridge, our cartridges are mechanisms for extending the capabilities of some underlying system. We differ in how our underlying system is implemented: data cartridges extend an Oracle server, but KG cartridges are implemented as and extend knowledge graphs while integrating with data analytics models.

HAO is inspired by several existing analytics-focused ontologies, including the Data Science Ontology (DSO) associated with the semantic flow graph [18] approach and the analytics ontology associated with the ScalaTion [15] framework. In semantic flow graphs, functions in an analysis script are mapped to abstract concepts defined in the DSO; graph visualization allows for language-independent workflow summarization. With ScalaTion, axioms and an analytics model taxonomy allow model selection to be performed via inference. In STA and HAO, we focus on the problem of domain-guided subpopulation-based health analysis in survey-weighted data [8]. Note that subpopulation discovery and representation require descriptive rather than predictive modeling workflows.

With a similar goal as Automated Machine Learning (AutoML, [22]), especially the Automatic Statistician [21], we aim to automate end-to-end statistical analysis. Our work differs from existing AutoML in several key areas. First, STA utilizes domain-dependent analysis techniques. The Automatic Statistician does not represent domain knowledge semantically; also, much of its work has been applied to nonparametric Bayesian models, such as Gaussian processes for time series [11]. In contrast, STA utilizes a variety of parametric statistical models, and we focus on the special case of data generated by a complex survey design. Unlike in a nonparametric model, the parameters of our models can act as explainable summaries of discovered associations. Finally, we note that the strategies of the Automatic Statistician or any other machine learning method can be readily incorporated into STA by representation in the KG.

2 Risk analysis in NHANES

Algorithm 1 illustrates how the STA framework uses semantic structures realized as cartridges to drive precision health subpopulation discovery and risk factor identification. In STAGE I, the STA framework queries its input cartridges to infer requirements for data preparation. This might entail filtering records for subjects that satisfy study inclusion criteria, log-transforming right-skewed variables, or constructing new variables based on supplied definitions. In STAGE II, an array of SCMs is trained on the prepared cohort using different hyperparameter configurations. A final model is determined by supplied model selection metrics, e.g., the Bayesian Information Criterion (BIC). In STAGE III, survey-weighted generalized linear models (GLMs) are trained on each discovered subpopulation. The regression coefficients and log-odds ratios estimated by these GLMs quantify the association between the supplied risk factor and response variable. After STAGE II and STAGE III, model findings and subpopulation characteristics are written to output cartridges for future reference.

In Algorithm 1, we perform precision risk analysis for a single risk factor. In practice, we repeat this process for many different categories of risk factors, yielding a precision environment-wide association study (EWAS, [16]). Similarly, the same risk factor can be tested against multiple potential response variables. Output cartridges generated by analyses are written back to the knowledge graph, where they are linked to the input cartridges used to generate them. This linkage grants STA explainability: all details of provenance and execution steps are captured, enabling detailed justifications for conclusions to be generated. Storing each piece of data and metadata in a knowledge graph also enables analysis reproducibility, since all details are kept together.

We present the STA framework for addressing population health problems. By varying models, variables, and the underlying datasets, we can adapt this workflow for other tasks. Classification and multiple regression models are used to identify potential risk factors – covariates that are strongly associated with the response variable in the study cohort, after controlling for known confounders.

NHANES is constructed with a multistage complex survey design (CSD, [8]) for each year. An NHANES subject's role in the CSD is captured by their survey weight, stratum, and variance unit; the STA framework encodes these values in the KG and then automatically utilizes them correctly in analyses. Since CSD data is not *iid*, incorporation of survey weights is necessary to attain unbiased statistical estimates. SCMs in STAGE II of Algorithm 1 use survey weights and Stage III uses the `survey` package in R, to create survey-weighted GLMs that incorporate the weights, strata, and variance units encoded in the KG.

3 Serialized components for automatic analysis

Minimal necessary analysis components from the HAO are stored in modular serializations called *cartridges*. A cartridge is a subgraph containing application- and analysis-specific entities. Cartridges can be edited to include additional elements, thereby enabling the flexibility to address a range of problems with

```
Input: NHANES subject records, response-cartridge, cohort-cartridge,
       risk-factor-cartridge, parameters-cartridge
Output: model-cartridge, results-cartridge, subpopulation-cartridge
STAGE I: DATA PREPARATION
· Select all subjects satisfying cohort-cartridge's inclusion criteria and store
  as cohort
· Query parameters-cartridge and risk-factor-cartridge for preprocessing
  techniques and apply to cohort
· Query response-cartridge and risk-factor-cartridge for necessary
  control-variables
· Query risk-factor-cartridge for risk-factor
· Query response-cartridge for response-variable
· Query parameters-cartridge for model-selection metric
· Calculate population-level summary statistics of cohort and write to
  subpopulation-cartridge
STAGE II: SUBPOPULATION DISCOVERY
for every hyperparameter configuration in parameters-cartridge do
    · Train SCM on cohort using control-variables and risk-factor to
      predict response-variable
    · Calculate SCM's metric value
end
· Identify SCM with optimal metric value and write its optimal hyperparameters
  and parameters to model-cartridge
· Take optimal SCM and write to model-cartridge, serialized as pickle
STAGE III: RISK MODELING
for every subpopulation discovered by SCM do
    · Select members of cohort belonging to subpopulation
    · Calculate summary statistics of subpopulation and write to
      subpopulation-cartridge
    · Train survey-weighted GLM on subpopulation using control-variables
      and risk-factor to predict response-variable
    · Extract risk-factor's $p$-value, regression coefficient, and regression
      coefficient standard error from GLM and write to results-cartridge
end
```

Algorithm 1: STA SCM analysis for subpopulation-specific or precision risk analysis of a single risk factor. Implemented variants include examining many potential risk factors in succession via an EWAS, as well as the addition of STAGE IV: REPORT GENERATION, in which the output cartridges are used to automatically create a report describing the findings. Reports use text, tables in the style of Table 3, and figures in the style of Fig. 2.

minimal modification. In Section 5, we demonstrate this versatility. In STA, cartridges are loaded and modified as the user constructs their risk analysis study.

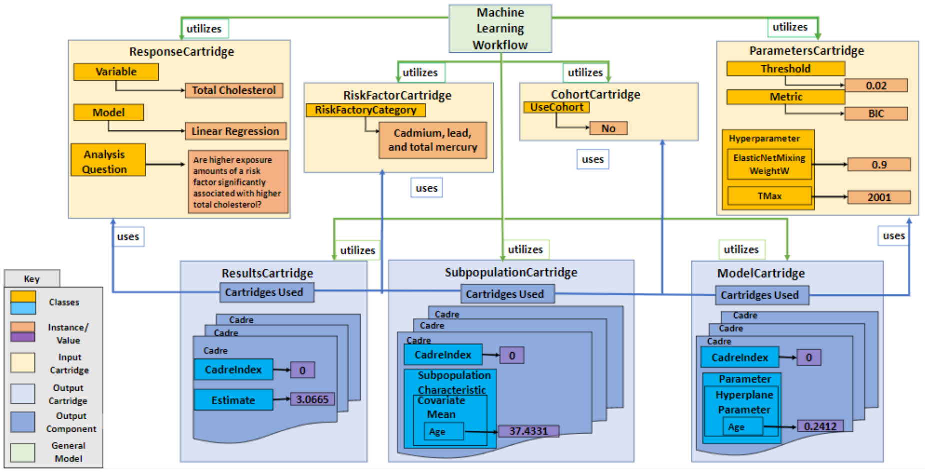

Fig. 1. Structural overview of the cartridge framework for a study using the SCM to identify subpopulation-specific associations between three heavy metals and total cholesterol. The descriptions of these cartridges are shown in Table 1.

3.1 Health Analysis Ontology

In the Health Analysis Ontology (HAO), we support modeling of processes, components, models, variables and factors involved in a health analysis pipeline such as the one described in Algorithm 1. The HAO reuses classes and properties from existing ontologies, listed in Table 2, but we also found it necessary to introduce new terminology.

We represent the ontology using OWL and introduce property associations between classes using owl:Restrictions. Overall, HAO provides a vocabulary necessary to model the reusable components of an analysis (sio:Analysis) implemented by an analysis workflow (hao:AnalysisWorkflow) that we store in cartridges (hao:Cartridge). Cartridges serve as containers that encode information about specific portions of a workflow. For example, a response cartridge (hao:ResponseCartridge) contributes to a high-level overview of a model with entities (modeled via sio:hasAttribute) such as the analysis question, response variable, type of model, etc. The HAO schema allows for the representation of cartridges as named knowledge graphs in the TriG format[3].

The HAO ontology only imports the SemanticScience Integrated Ontology (SIO), as we reuse several classes from SIO and utilize their object properties

[3] Learn more at https://www.w3.org/TR/trig/

to define associations between classes. For other terms that we reuse from large ontologies such as the National Cancer Institute Thesaurus (NCIT), the Statistical Methods Ontology (STATO) and the Ontology of Biological and Clinical Statistics (OBCS), we apply the Minimum Information to Reference an External Ontology Term (MIREOT, [4]) technique to include terms. HAO combines terminology from statistical, scientific and biomedical ontologies to model a reusable and modular health analysis pipeline. Additionally, to provide information on the intended usage of classes, we maintain metadata such as definitions (skos:definition) and descriptions (rdfs:description) on our ontology classes. We have tested the logical correctness of HAO by reasoning using the Hermit reasoner [20] in Protege [9]. The HAO ontology can be explored via online documentation[4] generated by using the Widoco [6] tool.

3.2 Cartridges

Cartridges can be grouped into two categories, input and output, with further subdivisions given in Table 1. Fig. 1 gives a high-level summary of the cartridge framework. In practice, cartridges are implemented as named graph collections (in TriG format) encapsulating instances of ontology classes that, when grouped together, represent different modules of an analysis workflow. Further, cartridges are constructed using terms from ontologies listed in Table 2. Domain-specific choices (e.g., choice of confounders or cohort inclusion criteria) about cartridge contents are adapted from published studies and linked with provenance. In the case that outdated or inaccurate knowledge is retired, this provenance shows what cartridges need to be updated.

Currently, input cartridges must be manually defined by domain specialists, but output cartridges are generated automatically after analysis. Minimal modification is needed to allow an input cartridge to be applied to a different analytics question. Cartridges can be edited to allow for the flexible tailoring of a health analysis pipeline to discover new subpopulations (stato:0000203 - cohort), identify new outcomes or test different response variables (hao:TargetVariable). For example, creating a new analysis of hypertension based on a type 2 diabetes analysis requires only a simple edit of the response cartridge; the other input cartridges remain the same. We maintain analysis related concepts in HAO, and for cartridges such as the subpopulation cartridge requiring domain-specific terminology we directly reference terms from ontologies in the field, within the cartridge. Additionally, as shown in Fig. 1, cartridges contain links to other cartridges that were used to generate it, to allow for easy traversal of all the components of a workflow.

4 Precision risk with supervised cadres

Our method for precision risk is the supervised cadre model [14], which simultaneously discovers subpopulations and learns their risk models. We use

[4] https://therensselaeridea.github.io/hao-ontology/WidocoDocumentation/doc/index-en.html

Cartridge category	Cartridge type	Contents
Input	Response	Analysis concepts and background domain axioms necessary to model a given health condition
	Cohort	Inclusion criteria used to determine if a given subject may be included in the user's study, which can be chosen on-the-fly or adapted from existing studies
	Risk factor	How categories of semantically-similar risk factors should be modeled
	Parameters	Rules to complete chosen analysis workflow and potential hyperparameter configurations for chosen model
Output	Model	The hyperparameters used to train a model, the parameter estimates learned during training, and the rules by which it is applied to new observations
	Subpopulation	Summary statistics characterizing discovered subpopulations, including within-subpopulation variable means and rates
	Results	Quantification of subpopulation-specific discovered associations between the risk factor and the response variable using regression coefficients, standard errors, and p-values

Table 1. Types of cartridges used in STA framework

Ontology	Prefix	Purpose
Health Analysis Ontology	hao	Inform analysis design, summarize analysis results for comparison, and generate reports
Study Cohort Ontology	sco	Represent cohort variables and control/intervention groups in Cohort Summary Tables of observational case studies and clinical trials
Children's Health Exposure Analysis Resource	chear	Represent the inclusion of environmental exposures in health research
The Statistical Methods Ontology	stato	Represent concepts and properties related to statistical methods and analysis
Semanticscience Integrated Ontology	sio	Provide an upper level ontology (types, relations) for consistent knowledge representation across physical, process and informational entities
National Cancer Institute Thesaurus	ncit	NCIT is an authoritative reference terminology in the cancer domain, but in our case we leverage its broad coverage and use it root to terminology on model-related parameters
Ontology for Biomedical Investigations	obi	Annotate biomedical investigations, including the study design, protocols used, the data generated and the types of analysis performed on the data
The PROV Ontology	prov	Model provenance information for different applications and domains
Ontology of Biological and Clinical Statistics	obcs	Represent additional biostatistics terms not in OBI
DC Terms	dct	Specify all metadata terms maintained by the Dublin Core Metadata Initiative
Simple Knowledge Organization System	skos	Define the new terms in the HAO

Table 2. Ontologies currently used in STA. The usage of these ontologies are described in sections 3.1 and 3.2

subpopulation-specific and precision interchangeably. The SCM is applied during STAGE II of Algorithm 1. Subpopulations, which we call cadres, are subsets of the population defined with respect to a cadre-assignment rule learned by the SCM. Subjects in the same cadre have the same association with a given risk factor. In STA, the chosen parameter and response cartridges set up the appropriate SCM and describe how to tune its hyperparameters. Optimal model parameters and hyperparameters are written to a model cartridge, which can be applied to novel subject records to determine their cadre.

We outline SCM for multivariate regression and binary classification. When trained on a set of subject records $\{x^n\} \subseteq \mathbb{R}^P$ and response values $\{y_n\}$, the SCM divides the observations into a set of M cadres. Each cadre m is characterized by a center $c^m \in \mathbb{R}^P$ and a linear regression function e_m parameterized by weights $w^m \in \mathbb{R}^P$ and a bias $w_m^0 \in \mathbb{R}$. New observations x have (for multivariate regression) an aggregate regression score (e.g., a subject's expected total cholesterol level) or (for binary classification) an aggregate risk score (e.g., the logit of their probability of having prediabetes) given by $f(x) = \sum_{m=1}^{M} g_m(x) e_m(x)$, where $g_m(x)$ is the probability x belongs to cadre m, and $e_m(x)$ is the regression or risk score for x were it known to belong to cadre m. These have the form $g_m(x) = \frac{e^{-\gamma||x-c^m||_d^2}}{\sum_{m'} e^{-\gamma||x-c^{m'}||_d^2}}$ and $e_m(x) = (w^m)^T x + w_m^0$.

Here, $||z||_d = \left(\sum_p |d_p|(z_p)^2\right)^{1/2}$ is a seminorm parameterized by $d \in \mathbb{R}^P$ and $\gamma > 0$ is a hyperparameter. SCM parameters are obtained by applying stochastic gradient descent to a survey-weighted loss function based on mean squared error or logistic loss, along with elastic net [24] regularization to improve interpretablity. The hyperparameters are chosen via a grid-search procedure and recorded in the chosen parameters cartridge. Compared to other nonlinear machine learning techniques, SCMs are more interpretable because of their within-subpopulation linearity. Examining the properties of each subpopulation and linear prediction model can yield significant insights. We have prototyped a system using the `shiny` R package that interacts with the user to design and conduct a study and then automatically generates interactive reports with text and figures explaining the results driven by the results cartridges and other external domain-specific linked-data. Sample results are presented in the next section.

5 Results

We present two risk analyses to identify subpopulation-specific environmental exposure factors associated with total cholesterol (TC) and prediabetic-or-worse glycohemoglobin levels (prediabetes). Elevated levels of serum lipids such as TC are recognized as risk factors for cardiovascular disease, and associations between TC and environmental exposure levels were identified previously [2, 23]. Other work also discovered associations between diabetes and environmental risk factors [16, 10]. Thus, it is worthwhile to identify subpopulation-specific risk factors associated with TC and prediabetes to improve health situations.

We chose a set of input cartridges shown in Table 3 for TC using control variables from prior studies [16, 23]. We extract 201 environmental exposure potential risk factors from NHANES 1999 to 2014 grouped into 17 classes such as phthalates (PHT) or polyaromatic hydrocarbons (PAH). Each class of potential risk factors has its own cartridge that describes its usage in analytics models. However, on the GitHub repository[5] we only host an example of the heavy metals risk factor cartridge used in this analysis. The number of survey subjects that have measurements for a given risk factor ranges from 1,406 to 15,218.

With our input cartridges, we run Algorithm 1 for every potential risk factor. Each risk factor is included in a single SCM that discovers subpopulations in the data. In STAGE II of Algorithm 1, each discovered subpopulation has its summary statistics written to a subpopulation cartridge to be stored in the KG. Characteristics of subpopulations with significant positive associations are visualized in Fig. 2A. In STAGE III of Algorithm 1, each subpopulation has a survey-weighted GLM trained on it, and the risk factor's regression coefficient and p-value are extracted. Due to the large number of hypothesis tests, false discovery correction is applied to these p-values before assessing significance α at a threshold specified in the study's parameters cartridge (here, $\alpha = 0.02$).

[5] Visit: https://github.com/TheRensselaerIDEA/hao-ontology

Cartridge type	Contents
Response	TC is a continuous response variable; subjects' age, Body Mass Index (BMI), Poverty Income Ratio (PIR), smoking habits, drinking habits, gender, marital status, and education level should be controlled for
Cohort	All available NHANES subjects
Risk factor	201 environmental exposure risk factors divided into 17 categories
Parameters	Standardize risk factor measurements; train models with $M = 1, 2$ and 3 cadres and choose best one using BIC for model selection; significance threshold of $\alpha = 0.02$ for GLM hypothesis tests

Table 3. Chosen input cartridges for TC risk study. The prediabetes risk study uses the same cohort, risk factor, and parameters cartridge, with a different response cartridge.

Fig. 2B shows significant regression coefficients for total cholesterol, for which most significant associations were on a population-level. TC's subpopulation-specific significant associations were total mercury (LBXTGH), blood lead (LBXBPB), perfluorodecanoic acid (LBXPFDE), and urinary lead (URXUPB), and prediabetes' subpopulation-specific sigificiant association was urinary lead (URXUPB). In Fig. 2A, subpopulations for which URXUBP and LBXTGH were risk factors for high TC have higher rates of being married and male and higher mean ages and BMIs; subpopulations for which LBXBPB, URXBPB, and LBXPFDE have higher rates of being divorced or living with a partner, and smoking and drinking. Similar figures can be shown for prediabetes risk factors and subpopulations, but they are omitted due to space constraints. One finding was the discovery of a subpopulation with a significant positive association between urinary cadmium (URXUCD) and prediabetes. Compared to the overall study cohort, subjects in this subpopulation had higher mean age and BMI and lower mean PIR.

Now that STA has discovered these associations, they can be used to motivate analyses that look for causal relationships, via, e.g., randomized control trials. The fact that subpopulations and associations are written back to the knowledge graph as cartridges, ensure that they are easily accessible by future queries, and guarantee study reproducibility.

6 Discussion

We have presented a semantically-targeted analytics framework via which risk factors specific to a subpopulation may be discovered in datasets. With the supervised cadre machine learning method, we simultaneously discover subpopulations and identify their significant risk factors. To support this we built a novel Health Analysis Ontology that captures analytics and health domain knowledge.

HAO and other ontologies provide structure for defining cartridges that are used for modular analysis pipelines. We leverage this semantic modeling to dynamically construct and execute a risk model and interpret results. Using STA, the system provides explainable insights for future population health studies in a scientifically rigorous and reproducible way.

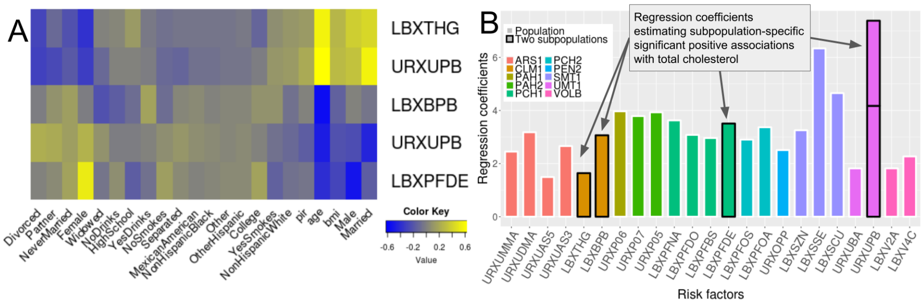

Fig. 2. Subpopulation characteristics and significant risk factors for TC study (A): Heatmap of normalized subpopulation-specific means for subpopulations with significant associations between a risk factor and TC. **(B):** Stacked bar chart showing significant positive associations with TC. Colors shows risk factor type.

In STA, statistical findings and parameters are encoded in results cartridges and written back to the KG, enabling retrieval for further study. Cartridges provide semantic extensions that enable a KG system to apply inference to solve domain-specific analytics problems. By publishing the results cartridges, studies become reproducible and explainable with provenance. Researchers with new analysis methods can readily compare results with prior studies, using the same workflow on the same problems. They can adapt existing peer-reviewed studies to new diseases by editing cartridge in published workflows.

We report here on semantically-targeted analytics applied to population health studies that rapidly enables new findings from the ongoing NHANES database. Using new cartridges, STA can readily be adapted to other types of statistical analysis on other data sources such as electronic health care records.

Acknowledgements

Created with support by IBM Research AI through the AI Horizons Network.

References

1. Al-Baltah, I.A., Ghani, A.A.A., Rahman, W.N.W.A., Atan, R.: A classification of semantic conflicts in heterogeneous web services at message level. Turkish Jnl of Electrical Engineering & Computer Sciences (2016)
2. Aminov, Z., Haase, R.F., Pavuk, M., Carpenter, D.O.: Analysis of the effects of exposure to polychlorinated biphenyls and chlorinated pesticides on serum lipid levels in residents of Anniston, Alabama. Environ Health **12**, 108 (Dec 2013)
3. Centers for Disease Control and Prevention (CDC): National Health and Nutrition Examination Survey (2017), http://www.cdc.gov/nchs/nhanes/
4. Courtot, M., Gibson, F., Lister, A.L., Malone, J., Schober, D., Brinkman, R.R., Ruttenberg, A.: Mireot: The minimum information to reference an external ontology term. Applied Ontology **6**(1), 23–33 (2011)

5. Frank, J.W.: Why "population health"? Canadian Journal of Public Health **86**(3), 162 (1995)
6. Garijo, D.: Widoco: a wizard for documenting ontologies. In: International Semantic Web Conference. pp. 94–102. Springer (2017)
7. Gietz, W.: What is a data cartridge? In: Data Cartridge Developer's Guide, chap. 1, pp. 1–17. Oracle Corporation (2002)
8. Heeringa, S., West, B., Berglund, P.: Applied Survey Data Analysis. Chapman and Hall/CRC., 2 edn. (2017)
9. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A.: The protégé owl plugin: An open development environment for semantic web applications. In: International Semantic Web Conference. pp. 229–243. Springer (2004)
10. Li, Y., Zhang, Y., Wang, W., Wu, Y.: Association of urinary cadmium with risk of diabetes: a meta-analysis. Environmental Science and Pollution Research **24**(11), 10083–10090 (Apr 2017)
11. Lloyd, J.R., Duvenaud, D., Grosse, R., Tenenbaum, J.B., Ghahramani, Z.: Automatic construction and natural-language description of nonparametric regression models. In: Proc. of the Twenty-Eighth AAAI Conf. pp. 1242–1250 (2014)
12. Martin, E., Monge, A., Duret, J.A., Gualandi, F., Peitsch, M.C., Pospisil, P.: Building an R&D chemical registration system. J Cheminform **4**(1), 11 (May 2012)
13. New, A., Bennett, K.P.: A precision environment-wide association study of hypertension via supervised cadre models. IEEE Journal of Biomedical and Health Informatics (2019), to appear
14. New, A., Breneman, C., Bennett, K.P.: Cadre modeling: Simultaneously discovering subpopulations and predictive models. In: 2018 Intl. Joint Conf. on Neural Networks (IJCNN). pp. 1–8 (July 2018)
15. Nural, M.V., Cotterell, M.E., Peng, H., Xie, R., Ma, P., Miller, J.A.: Automated Predictive Big Data Analytics Using Ontology Based Semantics. Int J Big Data **2**(2), 43–56 (Oct 2015)
16. Patel, C.J., Bhattacharya, J., Butte, A.J.: An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLOS ONE **5**(5), 1–10 (05 2010)
17. Pathak, J., Johnson, T.M., Chute, C.G.: Survey of modular ontology techniques and their applications in the biomedical domain. Integr Comput Aided Eng **16**(3), 225–242 (Aug 2009)
18. Patterson, E., Baldini, I., Mojsilovic, A., Varshney, K.R.: Teaching machines to understand data science code by semantic enrichment of dataflow graphs. CoRR **abs/1807.05691** (2018)
19. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web **8**, 489–508 (2017)
20. Shearer, R., Motik, B., Horrocks, I.: Hermit: A highly-efficient owl reasoner. In: Owled. vol. 432, p. 91 (2008)
21. Steinrucken, C., Smith, E., Janz, D., Lloyd, J., Ghahramani, Z.: The automatic statistician. In: Automatic Machine Learning: Methods, Systems, Challenges. pp. 175–188 (2018)
22. Yao, Q., et al.: Taking human out of learning applications: A survey on automated machine learning (2018), https://arxiv.org/abs/1810.13306
23. Zhang, S.H., et al.: Phthalate exposure and high blood pressure in adults: a cross-sectional study in China. Env. Sci. and Pollution Research **25**(16), 15934–15942 (Jun 2018)
24. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Stat. Society: Series B **67**(2), 301–320 (2005)