ICMLA 2011, Honolulu, HI

Using Published Medical Results and Non-homogenous Data in Rule Learning

Janusz Wojtusiak
George Mason University
Fairfax, VA, USA
jwojtusi@gmu.edu

Katherine Irvin
George Mason University
Fairfax, VA, USA
kirvin@gmu.edu

Aybike Birerdinc
INOVA Health System
Fairfax, VA, USA
abirerdi@yahoo.com

Ancha V. Baranova
George Mason University
Fairfax, VA
abaranov@gmu.edu

Abstract— **Many factors limit researchers from accessing studies' original data sets. As a result, much medical and healthcare research is based off of systematic reviews and meta-analysis of published results. However, when research involves the use of aggregated data from multiple studies, traditional machine learning-based means of analysis cannot be used. This paper describes diversity of data and results available in published manuscripts, and relates them to a rule learning method that can be applied to build classification and predictive models from such input. The method can be used to support meta-analysis and systematic reviews. Two application areas are used to illustrate the discussed issues: diagnosis of liver diseases in patients with metabolic syndrome, and detection of polycystic ovary syndrome.**

Keywords- Aggregated data, Meta-analysis, Published result, Rule learning, Systematic reviews

I. INTRODUCTION

Systematic reviews along with secondary data analysis are an important way of gaining medical and healthcare knowledge. They often require significantly less resources than performing randomized clinical trials. The goal of preforming systematic reviews is to bring together evidence from multiple published studies (often contradictory) and arrive at general conclusions. Large groups of people within the Cochrane Collaboration, Campbell Collaboration, and similar collaboratives use strict methodologies and guidelines in performing systematic reviews. This is because the original data used in studies are rarely available due to patient privacy, confidentiality, and reluctance for data sharing among research groups [17], and only the contents of the published work are available.

Despite the significant growth of machine learning, methods developed have not traditionally been used to support systematic reviews. One important reason for this is that the majority of machine learning algorithms are designed to work with data on an individual level (i.e. representing individual patients), but not with summaries, results, and aggregates in various forms (like those used in systematic reviews). Therefore, the analysis of published results and data requires a new specialized class of machine learning methods able to deal with non-homogenous inputs.

The goal of the presented research is to develop a methodology for building classification or predictive models from results and aggregated data present in published papers. It is sufficient that these papers contain aggregated data and/or results related to the problem.

This paper aims at presenting different forms of data within healthcare publications, and discusses challenges related to learning from them. It also briefly discusses modifications to the AQ21 rule learning system that enables learning from several of the discussed forms of data. Although the focus of this paper is on machine learning of attributional rules, and in particular the AQ21 system, most of the presented methods are applicable to other knowledge representations (decision trees, Bayesian networks, etc.).

Two applications of the AQ21 system on aggregated clinical data collected from studies on patients with metabolic syndrome (MS) and polycystic ovary syndrome (PCOS) are presented in this paper. The areas were chosen for both methodological and medical reasons. The methodological reason being that numerous publications are available in these two areas making it relatively easy to access. The medical reason being that not only do these conditions affect very large populations of patients, but there is no consensus on their causes and the best practices.

The central objective of this paper is to summarize types and forms of data that can be derived from publications, as well as present how machine learning can be used to analyze the data. It is not the intention of this paper to discuss how the data can be automatically derived from publications, nor how machine learning can help in selecting relevant publications for systematic reviews.

II. RESULTS AS DATA

One of the key challenges in using machine learning to analyze results is the diversity of the published studies. Publications include tables that summarize data, figures illustrating trends among cohorts of patients, correlation coefficients among selected (usually not all) variables, study inclusion criteria, definitions, and others. These data are very different from individual examples for which machine learning methods are designed to work with. This section, illustrated by MS and PCOS applications, briefly characterizes types of data/information/knowledge present in healthcare publications, and presents the process of using them to prepare the input data for machine learning.

A. Literature selection

Published literature is represented in a wide array of different research studies including clinical trials, observational studies, and comparative effectiveness studies. Having different types of research studies creates two major distinctions between the data obtained from the studies – the type of data (qualitative/quantitative or baseline/outcome) and the number of cohorts, participants, and data variables obtained from each. Additionally, it is important to consider how the data were collected (i.e., through randomized trial, observationally, or existing medical records) which affects the reliability and the biases within the data.

The selection of works to be used is a widely discussed issue [7], which applies to all systematic reviews and meta-analyses. It is usually conducted by a panel of experts and is based on criteria such as relevance of studies, quality of publications, cohort sizes, and others.

For example, of the twenty-five PCOS-related papers analyzed, eight of them could be classified as comparative effectiveness studies that compared drug treatments, tests, surgeries, environmental/genetic factors, or life style changes in correlation with PCOS. Another eleven of the papers could be identified as clinical trials. These contained the most significant amounts of data used in the analysis in terms of the number of patient cohorts and measured attributes. The majority of the papers included information identifying and addressing PCOS as well as the medical developments within treatment. However, six of the twenty-five papers were eliminated during the prepping of data for machine learning because they contained no relevant qualitative or quantitative variable values that could be used.

B. What is included in publications

There is no standardization in what data are included in publications and in what forms. Although some journals require structured manuscripts with specific sections, there are no standard formats in which data and results are presented. This is mainly because it is not possible to create one standard template that fits the wide range of topics covered by journals. The data and results generally come in tables, plots (that are usually equivalent to tables), and as a free text. Preparing them for machine learning-ready is time consuming, and requires expertise in the subject and knowledge of machine learning methods.

One challenge faced when applying machine learning to published data is in developing common representation of data and results. For example, there may be available aggregated data tables and inclusion criteria for twenty patient cohorts, two predictive models, and a small individual patient dataset. Each provides unique inputs to the learning problem and should not be disregarded. Results and data include quantitative and qualitative information - definitions, individual patient data, aggregated patient data, inclusion criteria, predictive/classification models, correlation coefficients, and statistical significances.

1) Definitions

Often there is no consensus on the definitions of symptoms and diseases. For example, there is no one consistently used definition of MS - it is usually defined as a combination of obesity, insulin-resistance, hypertension, elevated triglycerides, as well as a decreased level of high-density lipoprotein cholesterol (“good cholesterol”). Similar situation applies to PCOS. Specific definitions of syndromes,diseases, and their variations need to be included in data, as illustrated in Table 1 for the PCOS case.

Table 1: Excerpt of article description & PCOS definitions

Index			Definition of PCOS			Other Features of PCOS								
ID	Title	Year	oligo/anovulation	HA	PCO	O	IR	IGT	DM2	DYS	CVD	SA	M.IR	IN
ST1	1. Diagnosis and Mgmt of PCOS A Practical Guide	2006	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes
ST2	2. Polycycstic Ovarian Syndrome: Diagnosis and Management	2003	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes		Yes	
ST3	3. Effect of a low glycemic index compared with a conventional healthly diet on PCOS	2010	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes		Yes	
ST4	4. Levels of lipoprotein and homocysteine in non-obese and obese patients with PCOS	2005		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes	

HA - Hyperandrogenism, PCO - Polycystic Ovary, O – Obesity IR - Insulin resistance, IGT - Impaired Glucose Tolerance, DM2 - Diabetes Mellitus 2, DYS - Dyslipidemia, CVD - Cardiovascular Disease, SA - Sleep Apnea, M.IR - Menstrual Irregularity, IN - Infertility

2) Individual patient data

A majority of machine learning algorithms are designed to work with individual level data points. For the reasons described in the introduction, and additional ones discussed by Wojtusiak et al. [17], such data are often not available for multiple studies. In some situations, however, there may be limited individual datasets accessible for one or few studies, but not for others.

The most common form of individual data is attribute-value pairs. Because each patient is often described using the same set of attributes, the actual input data is in the form of a flat data table (it is important to note that some machine learning methods can deal with relational data [13], and predicates [6]). The usual situation is that each patient is an individual data point described using a set of attributes A_l … A_k. Typical machine learning programs use such individual data points in the form of attribute-value examples (1) where v_1, v_2, …, v_k are values of attributes A_1, … A_k.

$$(v_1, v_2, \dots, v_k) \tag{1}$$

Each example is described using the same attributes, thus the input dataset used for learning is in the form of a flat attribute-value table. In the case when some attributes are not present in a description of a specific example, meta-values (a.k.a. missing values) can be used. This form of data is, however, almost never included in published papers for reasons outlined in the introduction.

In the presented PCOS experimental results, no individual patient datum were used to create models (no such datum was available). Conversely, models for predicting MS liver complications were consequently tested for accuracy on an individual patients’ dataset.

3) *Aggregated patient data (tables and figures)*

Aggregated data are the most common way of describing cohorts of patients [16][17]. Aggregated values of attributes are given in the form of pairs (μ_A, σ_A), where A is a measured attribute, and μ_A and σ_A denote its mean and standard deviation measured over a group, for which the aggregation was done. Given that means and standard deviations for several parameters are available, each group can be described by an *aggregated example* given as (2).

$$((\mu_{A_1}, \sigma_{A_1}), (\mu_{A_2}, \sigma_{A_2}), \ldots, (\mu_{A_k}, \sigma_{A_k})) \qquad (2)$$

For non-numerical attributes, a typically used aggregated form lists the frequencies of values in a group, explicitly showing the distribution of examples. For example, a group of patients may include 70% White 17% Black, 2% Asian patients, and so on.

Aggregated data presents a significant challenge when using machine learning to analyze published results [16][17].

Table 2: Excerpt of aggregated data from PCOS study.

Identification			Study Variables (mean±std dev)														
ID	Cohort	Occurance	Subjects (n)	Age		Weight			BMI			C			T		
ST3	ST3CH1	Baseline	50	31	0.7	91.1	2.7	kg	34.3	1.0	kg/m²	4.8	0.1	mmol/L	1.3	0.1	mmol/L
	ST3CH2	Baseline	46	29.3	0.8	94.4	2.6	kg	34.7	0.9	kg/m²	4.8	0.1	mmol/L	1.2	0.1	mmol/L
	ST3CH1	Outcome	29			86.7	3.2	kg	32.7	1.2	kg/m²	4.7	0.2	mmol/L	1.3	0.1	mmol/L
	ST3CH2	Outcome	20			90.4	3.4	kg	33.2	1.2	kg/m²	4.5	0.2	mmol/L	1.1	0.2	mmol/L
ST4	ST4CH1	Baseline	85	23.2	4.92				25.82	3.44	kg/m²	169	31.4	mg/dl	103	34.4	mg/dl
	ST4CH2	Baseline	50	23.98	6.08				26.15	4.89	kg/m²	163	34.5	mg/dl	87.8	43.3	mg/dl
	ST4CH3	Baseline	38	23.51	6.18				31.55	5.78	kg/m²	186	53.3	mg/dl	126	76.5	mg/dl
	ST4CH4	Baseline	25	24.01	6.89				30.79	4.16	kg/m²	170	34.4	mg/dl	88.7	43.3	mg/dl
	ST4CH5	Baseline	47	22.95	5.49				20.99	2.14	kg/m²	155	30.8	mg/dl	85.4	20.8	mg/dl
	ST4CH6	Control	25	23.96	5.68				21.49	3.48	kg/m²	164	28.3	mg/dl	85.6	43.8	mg/dl
ST6	ST6CH1	Outcome	209	27.9	4				36	8.9	kg/m²						
	ST6CH2	Outcome	209	28.3	4				34.2	8.4	kg/m²						
	ST6CH3	Baseline	418	28.1	4.02	94.33	24.66	kg	35.23	8.66	kg/m²						

C – Cholesterol, T– Triglycerides

Of the twenty-five PCOS-related papers analyzed eighty-three different cohorts were identified. Seventy-two different qualitative and quantitative variables were identified and organized into one table, whose excerpt is presented in Table 2 above. The table contained qualitative data associated to study participant's characteristics and treatment. The table also contained quantitative data associated with study participants physiological and biochemical states.

4) *Inclusion criteria*

Another type of data that describes entire groups in medical or social publications is inclusion criteria for a specific study. It provides important facts about study's participants. Often exclusion criteria are provided instead of explicit inclusion criteria. Aggregated attributes, apply to all individual patients in the described groups. For, example if a study is performed among white males, then each individual subject in the data has precisely these values for attributes describing ethnicity and gender. This is in contrast to aggregated attributes outlined in the previous section. Table 3 depicts part of the inclusion criteria in the PCOS studies.

In contrast to traditional meta-analysis, in the presented method inclusion criteria do not need to match between studies, but rather need to be explicitly listed and are used as part of the input in the learning program.

Table 3: PCOD study inclusion/exclusion criteria excerpt.

ID	Study	Subjects (N)	# of Cohorts	Cohort Description	Inclusion Criteria												
					AH	HPA	CS	AAA	AOA	TD	KD	H	DM	CVD	OC	AD	LLM
ST8	w/ PCOS	61	2	Cohort 1 - M , Cohort 2 - CC	?	No	?	No	No	No	?	?	No	?	?	?	?
ST9	w/ & w/o PCOS	50	2	Cohort 1 - w/ PCOS, Cohort 2 - w/o PCOS	No	No	?	No	No	No	No	No	No	No	No	No	No
ST10	w/ & w/o PCOS	120	3	Cohort 1 - M +LOD, Cohort 2 - M +CC, Cohort 3 - M	?	?	?	?	?	?	?	No	No	?	?	?	?
ST12	Obese w/ PCOS	69	1	Cohort 1 - Baseline/Outcome	No	No	No	No	No	No	No	No	?	?	?	?	?
ST13	w/ & w/o PCOS	66	2	Cohort 1 - w/ PCOS, Cohort 2 - w/o PCOS	No	?	?	?	?	No	?	No	?	?	?	?	No

M – Metaformin, CC - Clomiphene Citrate, LOD - Laparoscopic Ovarian Drilling, P- Pioglitazone, AH - Adrenal Hyperplasia, HPA - Hyperprolactinemia, CS - Cushing's syndrome, AA - Acromegaly Adrenal Androgen-producing Tumors, AOA- Acromegaly Ovarian Androgen-producing Tumors, TD - Thyroid disorder, KD - Kidney Dysfunction, H - Hypertension, DM - Diabetes Mellitus, CVD - Cardiovascular Disease, OC - Oral Contraceptives, AD - Anti-Depressants, LLM - Lipid-Lowering Medication

5) *Predictive/classification models*

Published results may include existing classification or predictive models related to the learning task. Although complete models are rarely included, they may be obtained from authors (if not proprietary). Typically various kinds of regression models, decision trees, Bayesian networks, neural networks, and rule-based models are used.

Different approaches can be used in order to incorporate existing models: incremental learning, which requires existing models to be in the same form as the target model; sampling; using existing models to weight examples; and using existing models to label data in "interesting" cases (a form of active learning). In general, the problem of learning from existing models is not properly addressed in ML.

6) *Correlations*

Correlation coefficients, often reported in publications, can be used to better estimate quality (based on estimated coverage/accuracy) of a learned hypotheses. Virtually all methods for calculating quality of hypotheses (rules as well as other representations) are based on their coverage. When learning from published results, particularly from aggregated data, the numbers of individual patients satisfying candidate hypotheses (coverage) are not known, and need to be estimated. Additional information in the form of correlation coefficients can be used to better calculate the coverage.

Table 4: Excerpt of PCOD study correlation coefficients.

ID	Cohort	Cohort	Study Variables Correlation Coefficients																	
ST9	ST9CH1	ST9CH2	Preptin & Fasting Insulin						Preptin & HOMA-IR						Preptin & Ferriman-Gallwey Score					
			r	=	0.323	p	<	0.022	r	=	0.319	p	<	0.024	r	=	0.689	p	<	0.001
			FSH & Fasting Insulin						FSH & HOMA-IR						Adiponectin & FSH					
ST14	ST14CH1		r	=	0.41	p	=	0.01	r	=	0.33	p	=	0.03	r	=	0.4	p	=	0.01
	ST14CH3		r	=	0.59	p	=	0.01	r	=	0.54	p	=	0.01	r	=	0.67	p	=	0.005
ST16	ST16CH1		BMI & Adiponectin						HOMA-IR & Adiponectin						Insulin & Adiponectin					
			r	=	-0.654	p	<	0.005	r	=	-0.615	p	<	0.005	r	=	-0.639	p	<	0.005
ST23	ST23CH2-B	ST23CH2-O	Omentin-1 & BMI						Omentin-1 & WHR						Omentin-1 & Glucose					
			r	=	-0.4	p	=	0.072	r	=	-0.445	p	=	0.043	r	=	-0.471	p	=	0.031

7) *Statistical significance (p-values)*

Statistical significance usually given as p-values is another form of data present in publications. It can be used to weight data points and attributes when learning. Cohorts of patients and attributes from studies whose results are significant are more important for learning, thus are weighted higher.

Table 5: Excerpt of PCOD study statistical significance.

Identification			Study Variables								
ID	Cohort	Cohort	BMI	C	T	HDL-C	LDL-C	TE	SHBG	FG	FI
ST3	ST3CH1	ST3CH2	= 0.76	= 0.96	= 0.48	= 0.14	= 0.67	= 0.53	= 0.82	= 0.25	= 0.57
	ST3CH1	ST3CH2	= 0.61	= 0.43	= 0.51	= 0.48	= 0.99	= 0.41	= 0.67	= 0.6	= 0.74
	ST3CH1	ST3CH2	= 0.63	= 0.73	= 0.74	= 0.59	= 0.93	= 0.16	= 0.59	= 0.74	= 0.7
	ST3CH1	ST3CH2		0.61	= 0.63	= 0.68	= 0.94	= 0.18	= 0.53	= 0.71	= 0.77
ST7	ST7CH1	ST7CH2	= NS	= NS	< 0.01			< 0.01	< 0.01	< 0.01	= NS
	ST7CH3	ST7CH4	= NS	< 0.01	< 0.01			< 0.01	< 0.01	= NS	< 0.01
	ST7CH5	ST5CH6	= NS	= NS	= NS			< 0.05	= NS	< 0.05	= NS
ST8	ST8CH1	ST8CH2	= 0.13					= 0.12		= 0.24	= 0.7
ST18	ST18CH1	ST18CH2	= 0.1	= 0.9	= 0.2	= 0.2	= 0.8	< 0.001	= 0.001	= 0.7	= 0.05
	ST18CH3-B	ST18CH3-O	= 0.5	= 0.6	= 0.9	= 0.6	= 0.7	= 0.5	= 0.6	= 0.9	= 0.9
	ST18CH4-B	ST18CH4-O	= 0.8	= 0.2	= 0.1	= 0.4	= 0.05	= 0.02	= 0.4	= 0.3	= 0.01

C – Cholesterol, T– Triglycerides, TE – Testosterone, SHGB – Sex Hormone Binding Globulin, FG – Fasting Glucose, FI – Fasting Insulin

C. *Preparation of data for rule learning*

Before data can be prepared for input, it must be collected from the selected publications and organized into a relational database. The database consists of the aforementioned tables with definitions, inclusion criteria, aggregated patient data, correlation coefficients, and other statistical results (Table 6).

The majority of machine learning software, including AQ21 used in this study, takes input data in the form of data tables in which each example is represented using equal number of attributes. The major difference between AQ21 and other rule learning software is that it can use aggregated and individual examples. Because of the variance in what attributes are reported in different publications, missing values need to be extensively used (note the semantic distinction between unknown, not-applicable and irrelevant missing values described by Michalski and Wojtusiak [9]). For example, out of 152 attributes collected in the MS study, only one was reported in all publications. Similarly, in the PCOS study, about 78% of data values were missing – many of the attributes were present in only a single study.

Results are also reported in different units among the various studies. Appropriate conversions need to be done. Finally, all data are checked for consistency. This includes checking for normal or reasonable ranges of values, typos, and variable usefulness for the application at hand. An excerpt of the final data table used is shown in Table 6.

III. Rule Induction from Published Results

A. *AQ algorithm*

Many algorithms are available for inducing rules from individual data. Despite their differences, the algorithms have two common elements: rule construction, and rule evaluation. The described method for learning from published results and aggregated data focuses on these two algorithm elements, thus, it is applicable beyond the AQ-based systems described here. AQ is a family of rule learning systems, including the newest AQ21 [15], developed in the GMU Machine Learning and Inference Laboratory (www.mli.gmu.edu). AQ21's elements are designed specifically to deal with clinical data. In the core of all AQ systems is a simple version of A^q algorithm for generating attributional rules [8][15][17]. Attributional rules are more expressive than typical *IF…THEN…* rules because they may include a variety of constructs, including internal conjunction and disjunction, comparison of attributes, counting attributes, simple arithmetic expressions, exceptions, preconditions, and many others [8]. Rule learning usually results in more than one rule in the systems output (rulesets and ruleset families) – these rules usually should be interpreted together).

The AQ learning works in two main stages: rule construction and rule optimization. At the core of the first stage is a star generation algorithm, which creates multiple generalizations (in the form of attributional rules), called stars, of a selected positive example. These rules do not cover negative examples (except for noise). A combination of rules selected from one or more stars is used as a generated hypothesis. More detailed description of the AQ21 system is available in another ICMLA 2011 paper [18].

The rule generation stage can be extended to deal with aggregated examples and published results. This is done by operators that combine the logic of rule learning with statistical information derived from publications [16][17].

The most important issue within rule learning is the evaluation/estimation of coverage. Rule selection and optimization criteria are, to a large extent, based on how

Table 6: Except of PCOD standardized data ready for AQ21 rule learning.

Identification			Study Variables (mean±std dev)							
Cohort	Occurance	PCOS	Weight	BMI (kg/m²)	TE (ng/dL)	SHGB (ug/mL)	FG (mg/dL)	FI (ulU/mL)	FGS	E (pg/mL)
ST6CH1	Outcome	Yes	?	36 +/- 8.9	61.3 +/- 32	3.4 +/- 18.7	89.2 +/- 16.5	22.6 +/- 20.7	14.7 +/- 8.2	?
ST6CH2	Outcome	Yes	?	34.2 +/- 8.4	63.1 +/- 28.4	3.6 +/- 20.3	88.9 +/- 18.6	22.4 +/- 30	14.4 +/- 7.4	?
ST6CH3	Baseline	Yes	94.33 +/- 24.66	35.23 +/- 8.66	62.03 +/- 28.63	3.4 +/- 18.0	?	22.99 +/- 26.63	14.4 +/- 7.88	?
ST8CH1	Baseline	Yes	?	29.7 +/- 6.13	119 +/- 84	?	91.71 +/- 9.45	15.62 +/- 18.13	12.67 +/- 7.66	49.59 +/- 25.37
ST8CH2	Baseline	Yes	?	27.3 +/- 5.9	86 +/- 44	?	86.58 +/- 13.24	8.77 +/- 5.43	9.71 +/- 6.19	38.73 +/- 17.02
ST9CH1	Baseline	Yes	53.04 +/- 5.32	22.53 +/- 5.31	62.79 +/- 27.13	5.2 +/- 5.0	90.84 +/- 11.14	20.55 +/- 13.4	15.4 +/- 5.16	42.01 +/- 16
ST9CH2	Control	Yes	54.33 +/- 12.3	20.54 +/- 1.95	49.06 +/- 16.8	6.1 +/- 2.0	90.96 +/- 14.88	12.6 +/- 5.92	9 +/- 2.41	109 +/- 85.8
ST19CH1	Control	Yes	80.5 +/- 4.3	30 +/- 0.6	37.4 +/- 5.7	7.6 +/- 0.5	82.8 +/- 3.5	7.2 +/- 3	4.8 +/- 1.2	?
ST19CH2	Baseline	Yes	80.2 +/- 3.7	30.7 +/- 1.2	103.7 +/- 11.5	4.9 +/- 0.5	97 +/- 9.6	21.3 +/- 5.9	15 +/- 1.9	?
ST19CH2	Outcome	Yes	79.6 +/- 8.4	?	92.2 +/- 14.4	6.2 +/- 1.3	85.4 +/- 6.1	11.6 +/- 1.4	16 +/- 3	?
ST22CH1	Outcome	Yes	?	30.2 +/- 1.6	75 +/- 49	?	?	19.4 +/- 2.9	21.5 +/- 1.5	42.1 +/- 5.8
ST22CH1	Outcome	Yes	?	28.5 +/- 1.8	57.2 +/- 32	?	?	13.8 +/- 1.9	14.6 +/- 0.7	66.4 +/- 11.1

TE – Testosterone, SHGB – Sex Hormone Binding Globulin, FG – Fasting Glucose, FI – Fasting Insulin , E - Estradiol many positive and how many negative examples in the training/evaluation data satisfy the rules. The method used in AQ21 estimated rules' coverage of individual patients is based on published results included in the data. Information on the distribution of patients, correlation coefficients, and statistical significance is used to provide better estimation.

IV. Example Applications

A. *Metabolic Syndrome-related liver complications*

The goal of this application was to create models for diagnosing liver complications in metabolic syndrome (MS) patients. MS and its secondary complications pose a significant challenge for practicing diagnosticians. The syndrome is associated with a resistance to insulin, and predisposes patients to non-alcoholic fatty liver disease (NAFLD) and its more severe manifestation, nonalcoholic steatohepatitis (NASH). Currently, it is not possible to make an accurate diagnosis of NAFLD and/or NASH without a liver biopsy (an invasive and costly procedure with complications including the death as a result of bleeding or infection). Therefore, simple "rules" that can aid physicians at diagnosing NAFLD are of great importance.

Below we present two example rules derived from aggregated data by the previously described method. The first rule states that *there is presence of non-alcoholic fatty liver disease or its subtypes, if body-mass index is greater or equal 26.85, except for when aspartate aminotransferase level is less or equal 27.2 units/L and adiponectin level is at least 7.25 mg/ml.* Numbers following the conditions and rules represent numbers of patients or groups of patients supporting/contradicting the rules (see the AQ21 User's Guide for more annotation information [14]).

```
[Class=NAFLD]
  <== [BMI>=26.85: 8,2]
    |_ [AST<=27.2] &
       [Adiponectin>=7.25]
   : p=8,n=0,cx=25
```

```
[Class=NAFLD]
  <== [Adiponectin<=6.18: 8,1]
      :p=8,$n_{min}$=0,$n_{max}$=1,cx=5
```

Similar rules have been obtained for predicting simple steatosis and nonalcoholic steatohepatitis [16]. The rules are easy to interpret and are consistent with experts' existing knowledge. Additional testing validated these rules for predicting NAFLD and resulted in a positive predictive value (PPV) of 85-87%, reflecting relatively high "rule-in" characteristic of the algorithm. The best rule for the prediction of NASH has an accuracy of 78%, with PPV of 71%, and negative predictive value (NPV) of 37%.

B. *Diagnosing Polycystic Ovary Syndrome*

PCOS is an endocrinopathic disorder – a hormone condition effecting fertility - that impacts 5 to 10% of females ages 18 to 44 [4]. As a result, copious amounts of diverse aggregated data are available in medical journals. While there is no consensus the exact cause of PCOS (though genetic and environmental factors have been suggested), researchers have found that there is a strong correlation between PCOS and obesity, type 2 diabetes, and insulin resistance.

The initial application of the AQ21 system revealed several rules that can be used to diagnose PCOS. Because of the initial status of the research, we do not claim any medical significance of the rules and further analysis is needed.

```
[ PCOS = Yes ]
  <== [ Thyroid_Disorder = Yes]
        : p = 392, on = 58, cx = 7
 <== [ Testosterone >= 56.169552]
        : p = 304, on = 148, cx = 5
 <== [ DHEA_sulfate >= 2321.65]
     : p = 118, on = 122, cx = 5
 <== [ Free_Testosterone >= 2.285]
      : p = 19, cx = 5
```

```
[ PCOS = No ]
  <== [ HDL_C >= 40.830116]
      : p = 116, cx = 5
  <== [ BMI = 17.25..24.525 ]
      : p = 61, on = 134, cx = 7
   <== [ Thyroid_Disorder = Yes]
      : p = 48, cx = 7
  <== [ Testosterone <= 41.815734]
     & [SHBG >= 4.3076307]
      : p = 10, cx = 10
```

Left column lists selected rules for classifying patients as having PCOS and right column lists those as not having PCOS. Note that for some cases these sets of rules may not provide a definitive answer because the sets of rules are partially intersecting, and do not fully cover a complete patient space. This follows the idea that it is better to give no answer than an incorrect one that misleads a diagnostician.

V. Related Research

The problem of analyzing results of published studies is well known. Meta-analysis methods, often used in systematic reviews, calculate statistical descriptions that characterize data used in multiple studies. Extensive theory has been built on how to aggregate results of multiple studies and derive at statistically valid conclusions [7]. These methodologies are used in preparing systematic reviews such as those by the Cochrane Collaboration [1] in healthcare, and the Campbell Collaboration [3] in public policy and law.

Rule learning is the most popular method for discovering relations between variables in large databases. Possibly the most straight-forward approach to learning from published aggregated data is to create individual instances by sampling and then use them in the learning process. This approach, however, does not work well in rule learning because artificial sampling introduces random relationships between attributes that mislead learning algorithms [10].

Two machine learning areas that are closely related to the described method are statistical relational learning [13] and inductive logic programming [6]. Both areas deal with the more general problem of learning from datasets with complicated structures, rather than the specific problem of learning from published results. Also, recently Bayesian networks are also used to support meta-analysis [11][12]. Despite the extremely fast growth of machine learning, machine learning has been used in systematic reviews mainly to help researchers select the most appropriate studies [2], not to create models, not to develop models from results.

A related field called Literature-based discovery (LBD) concerns methods for identifying unknown relationships in data drawn from published results [5]. By bringing together results published in several papers, new relationships that were not considered in the original studies can be found. While the general framework of LBD is somewhat similar to the described rule learning method, its goal is to discover relationships, rather than build models.

VI. Discussion and Conclusion

Published medical studies include a wide range of data that is not properly utilized by traditional systematic reviews. The described methodology for analyzing published results is intended to complement currently used techniques in systematic reviews and meta-analyses. With rapidly changing clinical knowledge, automated methods with the ability to incrementally update knowledge may prove to be the needed method to keep reviews up to date. This is particularly important because evidence-based medicine requires clinicians to access the newest results.

Preliminary applications of the method in the two areas, has shown that it is capable of producing simple rules that can aid physicians in diagnostic tasks. Though the method seems promising, further work is needed to accurately test and implement the created rules in clinical practice.

While the research on the methodology is still in progress, the envisioned extension of the AQ21 system will allow for the use of all types of data and results mentioned in this paper. Currently, AQ21 supports individual-level and aggregated data, and weights of attributes and examples. The system is envisioned to be a platform for human-oriented machine learning in medical and healthcare applications.

Systematic investigation and testing of the described methods is beyond scope of the short conference paper. However, we are in the process of using Monte Carlo simulation methods to prepare large number of simulated datasets that can be used to test the methods in diverse number of situations. In addition to the experimental testing, theoretical investigation of the method is being prepared.

Acknowledgments

The presented research has been conducted in part in the GMU Machine Learning and Inference Laboratory whose activities are supported by the National Institute of Standards and Technology and Mason-Inova fund grants.

References

[1] The Cochrane Collaboration, The Cochrane Manual 4, 2008. [updated 14 August 2008].

[2] Cohen, A.M., Ambert, K., McDonagh, M., “Cross-Topic Learning for Work Prioritization in Systematic Review Creation and Update” *JAMIA*, 16, 5, 690-704, 2009.

[3] Davies, F., Boruch, R., “The Campbell Collaboration Does for Public Policy what Cochrane Does for Health,” BMJ, 323, 294-295, 2001.

[4] DHQP. "FAQs - Polycystic Ovary Syndrome (PCOS)." *Women's Health - PCOS Fact Sheet*, U.S. DHHS, Office on Women’s Health, 17 Mar. 2010. Web. 17 July 2011 <http://www.womenshealth.gov/publications/our-publications/fact-sheet/polycystic-ovary-syndrome.pdf>.

[5] Gordon, M., Lindsay, R.K., Fan, W., “Literature-Based Discovery on the World Wide Web,” ACM TOIT, 2(4), 261–275, 2002.

[6] Lavrac N., & Dzeroski, S., *Inductive Logic Programming: Techniques and Applications*, Ellis Horwood, NY, 1994.

[7] Lipsey M.W., Wilson, D., *Practical Meta-analysis*, Sage Publications, Inc., 2000.

[8] Michalski, R. S., "ATTRIBUTIONAL CALCULUS: A Logic and Representation Language for Natural Induction," *Reports of the MLI Laboratory*, MLI 04-2, George Mason University, Fairfax, VA, April, 2004.

[9] Michalski, R. S. & Wojtusiak, J., "Reasoning with Missing, Not-applicable and Irrelevant Meta-values in Concept Learning and Pattern Discovery," *Technical Report 2005-02*, Collaborative Research Center 637, University of Bremen, Germany, 2005.

[10] Michalski, R. S. & Wojtusiak, J., "The Distribution Approximation Approach to Learning from Aggregated Data," *Reports of the MLI Laboratory*, MLI 08-2, George Mason University, Fairfax, VA, 2008.

[11] Sciarretta, S., Palano, F., Tocci, G., Baldini, R., Volpe, M., “Antihypertensive Treatment and Development of Heart Failure in Hypertension: A Bayesian Network Meta-analysis of Studies in Patients With Hypertension and High Cardiovascular Risk,” *Archives of Internal Medicine*, 171(5), 384-394, 2011.

[12] Sutton, A.J., Abrams, K.R., “Bayesian methods in meta-analysis and evidence synthesis,” Statystical Methods in Medical Research, 10, 277-303, 2001.

[13] De Raedt, L., *Logical and Relational Learning*, Springer-Verlag, 2008.

[14] Wojtusiak, J., "AQ21 User's Guide," *Reports of the MLI Laboratory*, MLI 04-3, George Mason University, Fairfax, VA, September, 2004.

[15] Wojtusiak, J., Michalski, R. S., Kaufman, K. & Pietrzykowski, J., "The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and its Novel Features," The 18th IEEE ICTAI, Washington D.C., (2006).

[16] Wojtusiak, J., Michalski, R. S., Simanivanh, T. & Baranova, A. V., "Towards application of rule learning to the meta-analysis of clinical data: An example of the metabolic syndrome," *IJMI*, 78, 12, e104-e111, 2009.

[17] Wojtusiak, J., Baranova, A.V., “Model Learning from Published Aggregated Data” *Learning Structure and Schemas from Documents*, M. Biba and F. Xhafa (Eds.) Springer, 2011.

[18] Wojtusiak, J., Ngufor, C., Shiver, J., Ewald, R., “Rule-based Prediction of Medical Claims’ Payments: A Method and Initial Application to Medicaid Data” *ICMLA* 2011.