Identifying homogenous subgroups for individual patient meta-analysis based on Rough set theory

Eleazar Gil-Herrera, Athanasios Tsalatsanis, Ambuj Kumar, Rahul Mhaskar, Branko Miladinovic, Ali Yalcin, and Benjamin Djulbegovic

Abstract— **Failure to detect and manage heterogeneity between clinical trials included in meta-analysis may lead to misinterpretation of summary effect estimates. This may ultimately compromise the validity of the results of the meta-analysis. Typically, when heterogeneity between trials is detected, researchers use sensitivity or subgroup analysis to manage it. However, both methods fail to explain why heterogeneity existed in the first place. Here we propose a novel methodology that relies on Rough Set Theory (RST) to detect, explain, and manage the sources of heterogeneity applicable to meta-analysis performed on individual patient data (IPD). The method exploits the RST relations of discernibility and indiscernibility to create homogeneous groups of patients. We applied our methodology on a dataset of 1,111 patients enrolled in 9 randomized controlled trials studying the effect of two transplantation procedures in the management of hematologic malignancies. Our method was able to create three subgroups of patients with remarkably low statistical heterogeneity values (16.8%, 0% and 0% respectively). The proposed methodology has the potential to automatize and standardize the process of detecting and managing heterogeneity in IPD meta-analysis. Future work involves investigating the applications of the proposed methodology in analyzing treatment effects in patients belonging to different risk groups, which will ultimately assist in personalized healthcare decision making.**

I. Introduction

In medical research, meta-analysis is used to obtain pooled estimates of the treatment effects reported in various clinical research studies. The importance of meta-analysis stems from the necessity to combine research findings that if considered separately they would produce insignificant, non-generalizable, and unavailing results, unfit to inform medical practice. By systematically combining findings from similar studies it is possible to achieve the totality of evidence necessary to evaluate the efficacy of an investigated treatment.

The challenge researchers face when performing meta-analysis is how to integrate studies that present differences in the design, characteristics and reported effects. Such differences are formally acknowledged as heterogeneity and they are defined as any kind of variability among studies [1]. Typically, there are three types of heterogeneity found in meta-analyses: 1. *Methodological*, which refers to variability in the study design and risk of bias (e.g. randomization, allocation concealment, blindness etc.)[2, 3], 2. *Clinical*, which refers to the variability in the participants, interventions and outcomes studied (e.g. age, race, disease severity, disease progression, past treatment etc.) [2, 3], and 3. *Statistical*, which refers to variability in the observed outcomes [1, 3]. Failure to detect heterogeneity leads to misinterpretation of the summary effect estimates, which jeopardize the quality of the meta-analyses [2, 3] and may produce faulty estimations of the effects magnitude [4, 5]. Both methodological and clinical heterogeneity may result in statistical heterogeneity [6]. Researchers focus primarily on detecting statistical heterogeneity and subsequently on determining whether such heterogeneity is caused due to methodological or clinical variations between studies [1].

Assessing statistical heterogeneity relies on approaches that involve hypothesis testing [1, 7-9], such as the Chochrane's chi-square (Q) [10] and the I^2 measure [9, 11]. Higher values on these tests indicate high heterogeneity between studies. Both chi-square and I^2 tests focus on detecting heterogeneity yet are unable to identify the specific causes that underlie heterogeneity across studies [12]. The burden of explaining heterogeneity falls on the researcher.

To explore and explain the observed heterogeneity, meta-analysts conduct sensitivity analysis, based on the methodological quality of studies, and sub group analysis, based on a pre-specified trial or patient characteristics [3]. That is, the trials included in the meta-analysis are grouped according to pre-specified criteria. In case of individual patient data meta-analysis patients are grouped according to pre-specified clinical characteristics. However, these pre-specified criteria and clinical characteristics are generated in an ad-hoc manner and rely on the skills and medical knowledge of the researcher performing the meta-analysis. Thus, the results of meta-analysis may potentially differ depending on the experience of the meta-analyst.

Subgroup analysis [13] and meta-regression[14] are also applied to individual patient datasets (IPD) containing patient characteristics that may potentially influence the treatment effects. Determining which set of characteristics can be used to obtain homogeneous groups yields in a complex process, where subgroup analysis and meta-regression have been found prone to false positive results and ecological bias.

In this paper, we focus on meta-analyses of individual patient data and we propose a novel methodology to identify homogeneous groups of patients for managing the detected

This work was supported by the Department of Army under grant # W81 XWH-09-2-0175. The study was approved by the University of South Florida institutional review board (IRB # 100701).

E. Gil-Herrera A. Tsalatsanis, A. Kumar, R. Mhaskar, B. Miladinovic and B. Djulbegovic are with the Division of Evidence Based Medicine, Dept. of Internal Medicine, University of South Florida, Tampa, FL 33612, USA (e-mails: eleazar, atsalats, akumar1, rmhaskar, bmiladin, bdjulbeg @health.usf.edu)

A. Yalcin is with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA (e-mail: ayalcin@usf.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

heterogeneity. Our approach is based on Rough Set Theory (RST) [15] and has the potential to automatize the process of creating subgroups of patients with similar characteristics.

The mathematical principles that govern RST rely on the relations between objects. Using RST, we analyze and evaluate all possible relations between patients to obtain the minimum and dispensable information required to generate homogeneous subgroups of patients (i.e. patients with similar characteristics). We envision our methodology to operate in an automatic manner without the researcher intervention in selecting those characteristics that matter in grouping patients for meta-analyses.

II. Methodology

A. Dataset

Our dataset consists of individual patient data collected from nine randomized trials studying the effect of Allogeneic Peripheral Blood Stem-cell transplantation (PBSCT) compared to Bone Marrow transplantation (BMT) in the management of hematologic malignancies [16]. In total, 1,111 patients were enrolled. Records of 44 patients containing missing information were removed leaving the dataset with 1067 complete cases. Table 1 describes the details of our dataset.

B. Rough Set Theory

In RST, a dataset is represented by an information system defined as a pair $S = (U, A)$ where U is a non-empty finite set of objects that in our case represents the 1,111 patients. The set A represents a non-empty finite set of attributes called the condition attributes that corresponds to the characteristics of each patient. For every attribute $a \in A$, the function $U \rightarrow V_a$ makes a correspondence between an object (i.e. a patient) in U to an attribute value, which is called the value set of a. For example, from table 1, the value of the attribute "Age" can be 0, 1 or 2 for a given patient. A dataset including an outcome variable $d \notin A$, is termed as a decision system, defined as: $DS = (U, A \cup \{d\})$. The decision attribute in our data is the variable *"Death"* representing the overall survival of a patient given the characteristics described in A.

C. Indiscernibility and discernibility relations

Two objects (e.g. patients) $u, u' \in U$ are indiscernible with respect to a set of condition attributes $B \subseteq A$ if they have exactly the same values in all attributes, i.e: $a(u) = a(u') \, \forall \, a \in B$. This relation is called *indiscernibility relation* and is defined as:

$$IND(B) = \{(u, u') \in U^2 : \forall a \in B, a(u) = a(u')\} \quad \forall \, B \subseteq A \quad (1)$$

The *indiscernibility relation* captures the redundant information in the dataset. Every subset $B \subseteq A$, can be used for constructing this relation, however, only subsets that maintain the structure of the original dataset, i.e: $IND(B) = IND(A)$, are considered appropriate. Such a subset $B \subseteq A$, is termed as an exact reduct. In the case that it would not be possible to obtain an exact reduct, approximated reducts with acceptable quality of approximation are considered. The quality of approximation (α_B) of a reduct B quantifies the proportion of objects correctly allocated in a decision class by using only the attributes in B, i.e:

$$\alpha_B = \frac{|POS(B)|}{|U|} \quad (2)$$

where, $POS(B)$ is the set of all objects correctly assigned to the right decision class. In general, the higher the value of α_B, the more desirable the reduct is for constructing homogeneous subgroups.

Table 1. Dataset description

Variable	Description	Categories	%
Age	Patient age	0: <20	6.25 %
		1: (20,40]	47.82%
		2: (40, 65]	45.93%
Gender	Patient gender	1: Male	59.66%
		2: Female	40.34%
Diag	Diagnosis category	Acute lymphoblastic leukemia (ALL)	12.5%
		Acute myelogenous leukemia (AML)	33.52%
		Chronic lymphocytic leukemia (CLL)	0.28%
		Chronic myelogenous leukemia (CML)	43.47%
		Hodgkin's disease (HD)	0.09%
		Idiopathic myelofibrosis (IMF)	0.76%
		Myelodysplastic symdrome (MDS)	5.87%
		Multiple myeloma (MM)	1.04%
		Non-hodking lymphoma (NHL)	2.46%
StatTrans	Diagnosis status	0: Favorable (early-stage disease)	74.62%
		1: Unfavorable (late-stage disease)	25.38%
Mtx	Methotrexate for GVHD prophylaxis	1: Yes	43.84%
		0: No	56.15%
CondReg	Conditioning regimen used	1: Total body irradiation based (TBI)	41.19%
		2: Non TBI based	58.81%
GrowthFac	Use of post-transplantation growth factor	1: G-CSF	58.14%
		0: not used	41.85%
Alloc	Treatment	1: PBSCT	49.05%
		2: BMT	50.95%
Trial	Origin of the study	BR	5.30%
		US1	16.29%
		No	5.78%
		SA	5.10%
		FR	9.56%
		EBMT	30.21%
		CAN	20.36%
		US2	1.70%
		UK	3.69%
Death	Overall survival	0: Survive	59.75%
		1: Death	40.25%

On the other hand, the *discernibility* relation accounts for differences between objects in terms of their attribute values, i.e:

$$DIS_{DS}(B) = \{(u, u') \in U^2 : \exists a \in B, a(u) \neq a(u')\} \quad \forall \, B \in A \quad (3)$$

III. Identifying Homogeneous Subgroups in Individual Patient Dataset

We use the indiscernibility relation to build homogenous subgroups based on patients with the same characteristics and we use the discernibility relation to explore the characteristics that differentiate each subgroup. Fig. 1 depicts an overview of the proposed methodology, which is comprised of 4 processes: 1. Obtain reducts; 2. Create homogeneous groups; 3. Regroup based on similarities; and 4. Evaluate groups' heterogeneity.

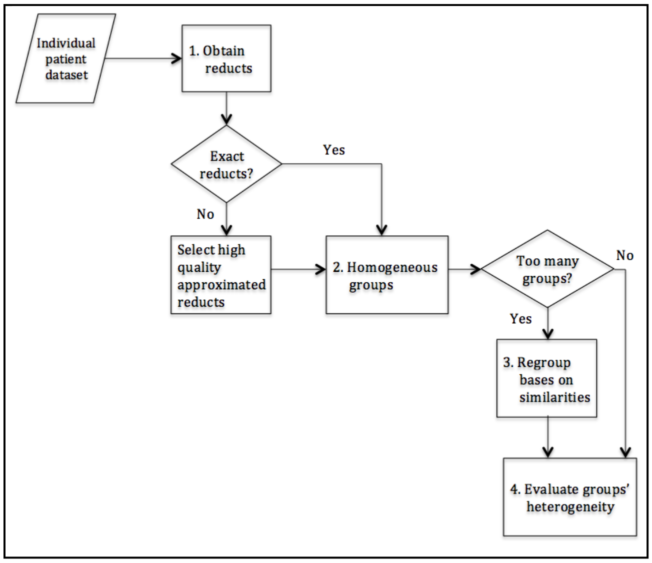

Individual patient dataset
1. Obtain reducts
Exact reducts?
Yes
No
Select high quality approximated reducts
2. Homogeneous groups
Too many groups?
No
Yes
3. Regroup bases on similarities
4. Evaluate groups' heterogeneity

Figure 1. Overview of the RST based methodology for identifying homogeneous subgroups in IPD

Obtaining reducts: First, we use the indiscernibility relation $IND(B)$ to obtain an appropriate subset of condition attributes B as the basis to generate the homogeneous subgroups of patients. To find this subset of attributes (reducts), we use approximated solutions described in [17]. In our dataset, the set $B = \{Age, Diag, StatTrans\}$ stands as the approximated reduct with the highest quality of approximation ($\alpha_B = 0.71$) among all the generated reducts.

Homogeneous groups: The indiscernibility relation partitions the IPD in 32 disjoint homogeneous subgroups with around 40% of them containing less than 10 patients. Subgroups with small number of patients do not include patients from all trials and are unsuitable for an individual patient meta-analysis.

Regrouping process: We obtain subgroups with a larger number of patients by merging smaller subgroups based on a similarity relationship. The similarity relation [18] is defined as a less rigorous version of the indiscernibility relation and is subject to a threshold value that allows small differences considered insignificant. Formally, we define the similarity relation between subgroups as:

$$g_1\, SIM_{B,\gamma}\, g_2\ iff\ \frac{|X|}{|B|} \geq \gamma\ , \forall\, g_1, g_2 \in U/IND(B)\ and\ u \in g_1 and\ u' \in g_2 \qquad (4)$$

Where, $X = \{a \in B : a(u) = a(u')\}$ and $\gamma \in [0,1]$ is the similarity threshold.

Since comparing all possible combinations between two groups to determine their similarity is a complex process we use a more straightforward procedure consisting in evaluating the differences between subgroups. Then, subgroups having similar differences to the rest of the subgroups are combined resulting in one homogenous group.

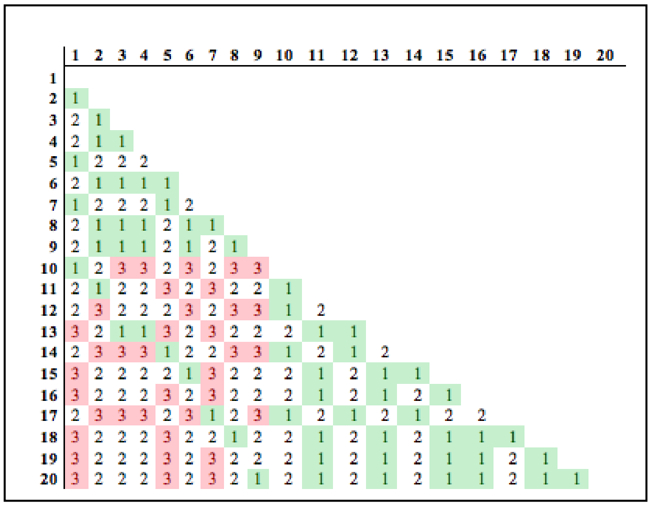
We define a discernibility matrix of subgroups $\mathcal{M}_B$, where each cell $\mathcal{M}_B(g_i, g_j)$ represents the number of attributes in B, whose values distinguish subgroup g_i from subgroup g_j, i.e:

$$\mathcal{M}_B(g_i, g_j) = \{|Dif|\}, where\ Dif = \{\, a \in B : a(u) \neq a(u')\}\ \forall\, g_1, g_2 \in U/IND(B)\ and\ u \in g_1 and\ u' \in g_2 \qquad (5)$$

Fig. 2 shows a portion of the discernibility matrix obtained for the 32 homogenous subgroups.

IV. Results

The initial 32 homogeneous subgroups are regrouped based on similarities in the number of attributes that distinguish them from the rest of groups. We chose a $\gamma = 0.8$ value (Equation 5) as a threshold parameter of similarity to minimize the number of homogeneous groups by allowing some degree of differences. For example, the initial subgroups 18, 19 and 20 (Fig. 2) can be regrouped since there are no more than 20% of differences across their corresponding rows. In other words, the three subgroups have similar distances, in terms of differences, to the rest of groups. As a result, the 32 homogeneous groups are gathered in three groups. Table 2 shows the homogenous groups resultant after the regrouping process. The mean number of patients in each group is equal to 355 with a standard deviation of 39.15.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2	1																			
3	2	1																		
4	2	1	1																	
5	1	2	2	2																
6	2	1	1	1	1															
7	1	2	2	2	1	2														
8	2	1	1	1	2	1	1													
9	2	1	1	1	2	1	2	1												
10	1	2	3	3	2	3	2	3	3											
11	2	1	2	2	3	2	3	2	2	1										
12	2	3	2	2	2	3	2	3	3	1	2									
13	3	2	1	1	3	2	3	2	2	2	1	1								
14	2	3	3	3	1	2	2	3	3	1	2	1	2							
15	3	2	2	2	2	1	3	2	2	2	1	2	1	1						
16	3	2	2	2	3	2	3	2	2	2	1	2	1	2	1					
17	2	3	3	3	2	3	1	2	3	1	2	1	2	1	2	2				
18	3	2	2	2	3	2	2	1	2	2	1	2	1	2	1	1	1			
19	3	2	2	2	3	2	3	2	2	2	1	2	1	2	1	1	2	1		
20	3	2	2	2	3	2	3	2	1	2	1	2	1	2	1	1	2	1	1	

Figure 2. A portion of the discernibility matrix obtained for the homogeneous groups. Each cell shows the number of attributes that differentiate between each pair of subgroups.

Table 2. Three homogeneous groups obtained from the regrouping process

Group number	Original group	Number of patients
1	10, 12, 14, 17	392
2	21, 23, 26, 29	359
3	1-9,11,13,15-16,18-20,22,24-25,27-28,30-32	314

The obtained homogeneous groups (Table 2) contain similar distributions in terms of trials, diagnosis and treatment. The statistical heterogeneity (I^2) indicate a negligible heterogeneity value for all the three groups (16.8% in group 1, 0% for group 2, and 0% for group 3), which suggests that all groups are indeed homogeneous.

V. Conclusions

In this preliminary work, we utilized a methodology typically found in engineering applications to solve a problem that exists in the realm of evidence-based medicine. Researchers who perform evidence synthesis are faced with the challenge of detecting heterogeneity between clinical trials and then explaining it by hypothesizing standards of similarity. However, there is no commonly accepted approach to identify similarities between trials and meta-analysts resolve to ad-hoc solutions. Here we presented a methodology based on Rough Set Theory that has the potential to automatize and standardize this process.

We demonstrated the effectiveness of our methodology using a sample dataset containing 1,111 patients from 9 different trials. We showed that were able to identify the appropriate patient characteristics to construct homogenous groups that presented similar proportion of trials, controls (diagnosis) and interventions (treatments) in accordance to the fundamental doctrine of meta-analysis. Thus, these groups are suitable to derive the pooled estimate of treatment effects in individual patient meta-analysis.

Other applications of this methodology include identifying subgroups of patients that need different treatments, patients with differential responses to therapy, or patients that belong to different risk groups. Analyzing the effect of treatment in each subgroup is very important for personalized healthcare. Our intention is to compare this methodology with similar approaches in other data sets.

Finally, this is a preliminary work and presents limitations. Particularly, we have not investigated the effects of our methodology in the results of meta-analysis, which we intent to do in the future. Other future research includes generalization of our methodology to accommodate clinical trial data in addition to individual patient data.

References

[1] J. P. T. Higgins, S. Green, and C. Collaboration, *Cochrane handbook for systematic reviews of interventions* vol. 5: Wiley Online Library, 2008.

[2] S. L. West, G. Gartlehner, A. J. Mansfield, C. Poole, E. Tant, N. Lenfestey, L. J. Lux, J. Amoozegar, S. C. Morton, and T. C. Carey, "Comparative effectiveness review methods: clinical heterogeneity," 2010.

[3] J. Higgins and S. Green. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. Available: http://www.cochrane.org/resources/handbook/index.htm

[4] K. Schulz, "Meta-analyses of interventional trials done in populations with different risks," *Lancet,* vol. 345, p. 1304, 1995.

[5] E. M. Balk, P. A. L. Bonis, H. Moskowitz, C. H. Schmid, J. P. A. Ioannidis, C. Wang, and J. Lau, "Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials," *JAMA: the journal of the American Medical Association,* vol. 287, pp. 2973-2982, 2002.

[6] S. G. Thompson, "Why sources of heterogeneity in meta-analysis should be investigated.," *BMJ,* vol. 19, pp. 1351-1355, 1994.

[7] J. P. Ioannidis, "Interpretation of tests of heterogeneity and bias in meta-analysis," *Jounral of Evaluation in Clinical Practice,* vol. 14, pp. 951-957, 2008.

[8] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, "Identifying and Quantifying Heterogeneity," in *Introduction to Meta-Analysis*, ed: John Wiley & Sons, Ltd, 2009, pp. 107-125.

[9] J. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Statistics in medicine,* vol. 21, pp. 1539-1558, 2002.

[10] W. G. Cochran, "Some Methods for Strengthening the Common $\chi 2$ Tests," *Biometrics,* vol. 10, pp. 417-451, 1954.

[11] J. Higgins, S. Thompson, J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *BMJ,* pp. 327-357, 2003.

[12] S. G. Thompson and S. J. Sharp, "Explaining heterogeneity in meta-analysis: a comparison of methods," *Statistics in medicine,* vol. 18, pp. 2693-2708, 1999.

[13] S. W. Wang R Fau - Lagakos, J. H. Lagakos Sw Fau - Ware, D. J. Ware Jh Fau - Hunter, J. M. Hunter Dj Fau - Drazen, and J. M. Drazen, "Statistics in medicine--reporting of subgroup analyses in clinical trials.", *New England Journal of Medicine,* vol. 357, pp. 2189 - 2194, 2007.

[14] J. Berlin Ja Fau - Santanna, C. H. Santanna J Fau - Schmid, L. A. Schmid Ch Fau - Szczech, H. I. Szczech La Fau - Feldman, and H. I. Feldman, "Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head.", *Statistics in Medicine,* vol. 21, pp. 371-87, 2002

[15] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Norwell, MA, 1992.

[16] H. Florida, "Allogeneic peripheral blood stem-cell compared with bone marrow transplantation in the management of hematologic malignancies: an individual patient data meta-analysis of nine randomized trials," *J Clin Oncol,* vol. 23, pp. 5074-5087, 2005.

[17] S. Vinterbo and A. Øhrn, "Minimal approximate hitting sets and rule templates," *International Journal of Approximate Reasoning,* vol. 25, pp. 123-143, 2000.

[18] R. Slowinski and D. Vanderpooten, "Similarity relation as a basis for rough approximations," 1995.