## An Interactive Data Extraction System to Create the Living Systematic Reviews and Meta-Analysis

## Irbaz Bin Riaz, MD, MS<sup>1,4\*</sup>, Huan He, PhD<sup>2\*</sup>, Syed Arsalan Ahmed Naqvi, MBBS<sup>3</sup> Rabbia Siddiqi, MBBS<sup>3</sup>, Noureen Asghar MBBS<sup>3</sup>, M. Hassan Murad, MD, MPH<sup>4</sup>, Hongfang Liu, PhD<sup>2</sup> <sup>1</sup>Department of Oncology, Mayo Clinic, Phoenix, AZ, USA <sup>2</sup> Department of AI and Informatics Research, Mayo Clinic, Phoenix, AZ, USA <sup>3</sup> Dow University of Heath Sciences, Karachi, Pakistan <sup>4</sup> Mayo Clinic Evidence Based Practice Center, Mayo Clinic, Rochester, MN, USA

**Introduction:** It takes months to years for conducting rigorous systematic review (SR) and meta-analysis (MA), and it involves developing a search strategy, screening for relevant citations, extracting and analyzing data and ironically the review can be outdated as soon as it published. The current process of creating SRMAs is inadequate to keep pace with rapid influx of evidence as seen in COVID-19 pandemic and dynamic fields like Oncology. Thus, living SRs (LSRs)—which are updated as soon as new evidence becomes available— are necessary to overcome the limitations of conventional reviews. Previously, we have described semi-automating the screening process and analyses, and here we describe our innovative method for data extraction to decrease the effort for creating and maintain LSRs.

**Methods:** We designed three major modules to facilitate the data extraction: the outline module (Fig. 1(A)), the table module (Fig. 1(B)), and the interactive extractor (Fig. 1(C)). The purpose of outline module is to create a skeleton of tables for data extraction. In the outline module, the users specify the outcomes of interest as well as the structure of the data format required to analyze each outcome. Users are provided with a range of standard options to create summary and data tables and are also provided with the flexibility to generate new variables as necessary. The tables in the table module are automatically populated with studies to be included in the review with the meta-data such as year of publication, author and journal name and rest of tables are completed with the help of extractor module. The studies are listed row by row. Users could select the "check" option and select each particular study for extraction. When a study row in this table is clicked, the interactive extractor would show its text details, including the abstract and full-text PDF files. The extractor module allows user to extract data from associated text. The user simply needs to highlight the text, right-click on the highlighted text, select which attribute the highlighted text belongs to. Then the highlighted text would be saved as the value of the selected attribute for this study. Thus, the tables generated from these modules can be presented to summarize the results of the systematic review or perform further analyses such as pairwise or network meta-analyses, create summary of findings tables and evidence maps.

E POPULATION CHARACTERI	Add Attribute		$(\Lambda)$	TRIAL CHARACTERISTICS	Very C Full 0	ub or Abstract 🔽 Dha	se Occierent/Sellen/Up ON		Abstract PDFs		$(\mathbf{C})$
Ø N	[ Treatment Arm	Control Arm		Treatment and Control Treatment Am							
Median Follow Un Durat	tion + -			SYS						1 1/04	
0.000				Is study included in Meta-analysis Is OS data available Is QoL data available The number of arms Risk URL				farms 🗆 Risk 🗆 URL			
S Median Pollow Up Duran				POPULATION CHARACTERISTICS					Masao Yamamoto Ramos, MD; Beatriz Mota-Vega, MD; Amir Carmona, MD; Marco Polo Peralta Alvarez, MD, MSc; Volanda Bautista, MD; Famando Aldaro, MD; Raquel Carson, MD; Christian Polo, MD, DhD, MBA, Rafael Posell, MD, DhD		
⊗ Sex	[ Males(%)	Females(%)	)+-	N [ Treatment Arm Control Arm ] Median Follow Up Duration (mo) Median Follow Up Duration for QoL (mo)					Totalida bautista, INIZ, Perhando Ardacti, MD, Raquei	se son, mo, christian kono, mo, Pho, more kalael kosel, mi,	A PHD
Risk Category Criteria	+-			Sex ( Males(%) = Females(%) ] = Bisk Category Included = HrQoL Instrument							
Risk Category Included +- PRO Completion Rate (%) [] BaselineOn Rx ] Ris						Risk Category (%) [   Favorable   Intermediate   Poor ]			IMPORTANCE Because of socioeconomic factors, many patients with advanced non-small cell		El Supplemental content
Hr0ol Instrument	+-			Oreans Involved at Reseline (%) [Clumph Node Cliner Clump Adrenal Cliner Cliner 1					lung cancer (NSCLC) do not receive immunotherapy in the first-line setting. It is unknown if		- sopprinting content
O page consultation from th	a needee	0.0		RESULTS FOR OVERALL POPULATION the control of the c					the combination of immunotherapy with ch	the combination of immunotherapy with chemotherapy can provide clinical benefits in	
O PRO Completion Rate (	(o) [ baseline	UNIEX	14-	Primary Endpoint(s) Secondary Endpoint(s) OS (mo) [ Rx Control ] PFS (mo) [ Rx Control ]							
Risk Category (%)	[ Favorable	Intermediate		ORR (%) Rx [ Overall CR PR ] ORR (%) Control [ Overall CR PR ] All-Cause Grade 3 or Higher (%) [ Rx Control ]							
Poor ]+- TRAE Grade 3 or Higher (%) [□Rx □Control ] TRAE Grade 5 (%) [□Rx □Control ] No of Deaths [□Rx □Control ] OBJECTIVE T									OBJECTIVE To evaluate the safety and effica	cy of the combination of pembrolizumab plus	
PD-L1 Positivity (%) [ Positive Negative ]+-  Significant QoL Benefit for Rx							docetaxel in patients with previously treated advanced NSLLL following platinum-based chemotherany regardless of FGEP variants or programmed cell death ligand 1 status				
Nephrectomy (%)	[ Treatment	Control	) <b>+</b> -	RESULTS BY RISK CATEGORY	RESULTS BY RISK CATEGORY					cicinite acting regime control management of programmed control and might an action	
Organs Involved at Base	lin f Lymph Node	Liver		Favorable MSK PFS [KK (mo)Control (mo) ] Favorable MSK ORR KR [OVerallCRPR ]					DESIGN, SETTING, AND PARTICIPANTS The Pembrolizumab Plus Docetaxel for Advanced		
Lung Adrena	al Bone	Brain	1+-	remotable Risk URK Control LOVerall Cox DV j Intervoor Risk Pris LOR (Intervoor Risk Pris LOR (I							
					Sen Orn Jr		dello della Oct Orit j		docetaxel alone from December 2016 throu		
	C Add their an			NCT Paper (35 selec	ed) Checked	Arms Phase	Treatment Arm	Control Arm	and the second sec	Phase	
E RESULTS FOR OVERVICE FO	El Auto Attribute			NCT01984242 12 - Pal SK 2020		2 - 2	bevacizumab	Sunitinib	pombrolizumah op day 8 (200 mg) overal 2	POPULATION CHARACTERISTICS	
Primary Endpoint(s)	+-			NCT02231749 178 - Robert J Mot:	er 🗆	2 🖌 3	Nivolumab plus ipil	Sunitinib	pembrolizumab maintenance until progress	N - Treatment Arm	
Secondary Endpoint(s)	+-			NCT01030783 179 - Robert J Mot.	er 🗆	2 🗸 3	Tivozanib	Sunitinib	received docetaxel monotherapy.	N - Control Arm	
@ 0S (mo)	[ Rx	Control	1+-	NCT01835158 180 - Toni K Choue	ri 2 🗆	2 🗸 2	Cabozantinib	Sunitinib	MAIN OUTCOMES AND MEASURES The prime	ry and point was overall response rate (OPP)	
PFS (mo)	f Bx	Control	1+-	NCT02853331 170 - Brian I Rini 20 NCT02853331 170 - Brian I Rini 20	19 🗆	2 - 3	Pembrolizumab plu	Sunitinib	Secondary end points included progression	free survival (PES), overall survival, and safety.	
Q 000 (%) Pr	( Ormali	CD.		NCT02684006 1/1 - MoDert J Mot	er U	2 • 3	Avelumab plus Axiti	Sunitinib	DESULTE Among 20 securited actionts 23 (	10/2	
PR 1+-	[ Overall	CR		NCT00420821 14 - Abons MB 2020		2 4 3	avitioib	Sunitinib	and 25 (32%) had an EGER/ALK alteration. E	orty patients were allocated to receive	
0.000 (0.000)	1.0			NCT02420821 164 - Rini B.I. 2019		2 - 3	Atezolizumab plus t	Sunitinib	pembrolizumab plus docetaxel, and 38 wer	e allocated to receive docetaxel. A statistically	
ORR (%) Control	[ Overall	CR		NCT00410124 161 - Motzer R.J. 20	08	2 - 3	sorafenib	Sunitinib	significant difference in ORR, assessed by a	n independent reviewer, was found in patients	
				NCT12345555 92 - Mennitto A 202	1 0	2 - 2	sorafenib	Sunitinib	receiving pembrolizumab plus docetaxel vs	patients receiving docetaxel (42.5% vs 15.8%;	
Grade 3 or Higher Aes (9	N) [ Rx	Control	1+-	NCT02574591 81 - Arrieta 0. 202	, 🗆	2 🛩 2		Sunitinib	odds ratio, 3.94; 95% CI, 1.34-11.54; P = .01)	Patients without EGFR variations had a	*
No of Deaths	( Rx	Control	)+-	NCT12345678 80 - Kang C. 2020		2 🛩 3	Atezolizumab	Sunitinib	· ·		•

Figure 1. The screenshots of our extraction tool. (A) the outline module; (B) the table module; (C) the interactive extractor.

**Discussion:** Extracting data for LSRs is particularly challenging as it requires management of the key information from meta data and unstructured free texts such as PDF files and web pages. Although existing tools, such Rayyan and Covidence, work well on independent tasks (e.g., screening, data retravel) and generate data files, linking those isolated data files correctly is challenging and requires tedious manual operations in multiple tools. Thus, to ease the burden of using we created a user-friendly interface which brings together data from multiple sources and facilitates data extraction and creation of summary tables and data tables for analysis. Next steps include integrating this data extraction system with our previously created modules for screening and analyzing the data thereby creating a pipeline for true, living systematic reviews which will be updated in "almost" real time.

\* Those authors have contributed equally as co-first authors