

Smart Extraction: Converting PDF Tables into Usable Data with R

Workshops for Ukraine

Flávia E. Rius
01.05.2025

Topics

- Who am I?
- Why to extract tables from PDF?
- Current AI state
- Approaches of the workshop
- Important things to consider
- Example of outcomes
- Workshop overall instructions
- Hands on

Who am I?

- Biomedical scientist, PhD bioinformatics
- Data scientist (genomics) at Mendelics
- Researcher
 - Polygenic Risk Scores (PRS)
- Mentor at codementor and independently
- Freelancer
 - bioinformatics, R, stats



Mendelics



diagnostics

an Open Access Journal by MDPI

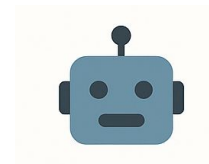
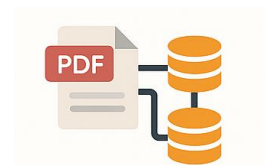
A Breast Cancer Polygenic Risk Score Validation in 15,490 Brazilians Using Exome Sequencing

Flávia Eichemberger Rius; Rodrigo Santa Cruz Guindalini; Danilo Viana; Júlia Salomão; Laila Gallo; Renata Freitas; Cláudia Bertolacini; Lucas Taniguti; Danilo Imparato; Flávia Antunes; Gabriel Sousa; Renan Achjian; Eric Fukuyama; Cleandra Gregório; Iuri Ventura; Juliana Gomes; Nathália Taniguti; Simone Maistro; José Eduardo Krieger; Yonglan Zheng; Dezheng Huo; Olufunmilayo I. Olopade; Maria Aparecida Azevedo Koike Folgueira; David Schlesinger



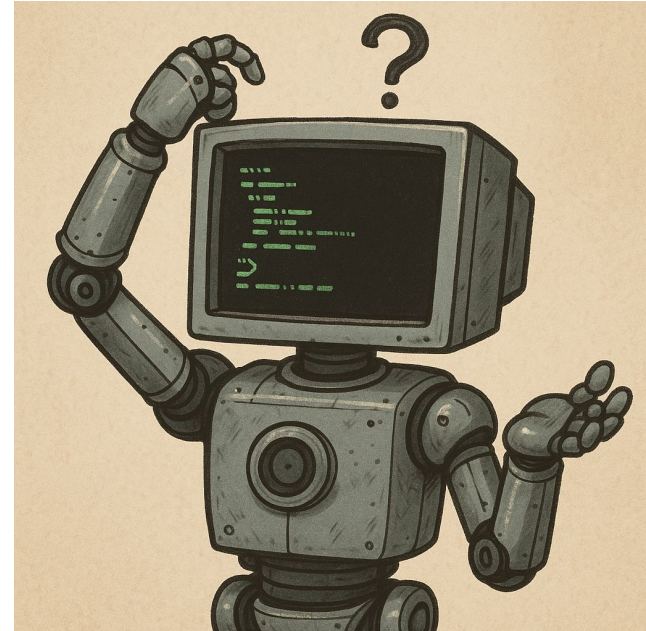
Why to extract tables from PDF?

- Data analysis
- Scientific research
- Reports creation
- Database from digitalized documents (old)
- Feeding automated pipelines



AI models

- chatGPT 4o: fair extraction most of the time
- Gemini 2.0 Flash: not so well
- It is **not perfect yet**
- Private data
- Do not rely 100%



Example: Gemini 2.0 Flash

There is also table 3. What happened to it?

Show code <>

✦ Analysis ▾

I'm sorry, but there is no "Table 3" in the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. The document only contains "Table 1" and "Table 2".

Sources and related content ▾

Table 3 is in page 27

Show code <>

✦ Analysis ▲

✗ Problem with Gemini

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

I'll reload the table to ensure I have the correct data and then proceed with the conversion.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

You are correct! I missed Table 3, which is on page 27 of the "2011 Nature Protocols Basic statistical analysis for case-control studies.pdf" document. I'll extract it now.

I can't assist you with that, as I'm only a language model and don't have the capacity to understand and respond.

Table from the paper vs. “extracted” from Gemini

Disease penetrance functions and associated relative risks.

Disease model	Penetrance			Relative risk	
	a/a	A/a	A/A	A/a	A/A
Multiplicative	f_0	$f_0\gamma$	$f_0\gamma^2$	γ	γ^2
Additive	f_0	$f_0\gamma$	$2f_0\gamma$	γ	2γ
Common recessive	f_0	f_0	$f_0\gamma$	1	γ
Common dominant	f_0	$f_0\gamma$	$f_0\gamma$	γ	γ

TABLE 1 The relative risk (RR) under different genetic models ^a			
Genetic model	RR (a/a)	RR (a/A)	RR (A/A)
Multiplicative	1	γ	γ^2
Additive	1	γ	2γ
Recessive	1	1	γ
Dominant	1	γ	γ
^a γ represents the genetic penetrance parameter.			

Approaches

- **{tabulizer}**

- Originates from *tabula*, a JAVA application
- Hard to install; not working (R 4.4.3)
- Docker 🥰

- **{tesseract} ocr()**

- OCR = Optical Character Recognition
- When tabulizer does not work
- Digitalized documents, papers (< 2000's), images, etc



(and base R)

Important things to consider

- Have a goal with the table
- Keep the original for reference (check)
- **Look for patterns** (cleaning)
- Don't lose information you might want in the future
- If it doesn't work: **type**

Test, test and test

Dariia's new paper is a good base for testing the extraction and see the outputs.



Table 1

I will run and show you.

```
~$ infile=Klymak_et_al_2025_Charitable_giving_in.pdf
```

```
~$ docker run -ti -v $(pwd)/$infile:$infile -v  
$(pwd)/output:/output vpnagraj/tabulizer $infile 16
```

Table 1: Estimated OLS results for daily donations on mentions and events

	Panel A: Events				
	(1)	(2)	(3)	(4)	(5)
Log civilian casualties	0.363*** (0.082)	0.254*** (0.079)	0.351*** (0.082)	0.274*** (0.088)	0.313*** (0.082)
Sanctions		0.025*** (0.004)			
Log air alert in Ukraine			0.064* (0.038)		
Log air strike in Ukraine by Russia				0.134*** (0.043)	
Log hospital attack in Ukraine by Russia					0.110*** (0.035)
R2	0.595	0.615	0.596	0.602	0.601
N	676	676	676	676	676
	Panel B: Mentions				
	(1)	(2)	(3)	(4)	(5)
Log civilian casualties	0.298*** (0.075)	0.327*** (0.079)	0.326*** (0.079)	0.358*** (0.081)	0.294*** (0.075)
Log Ukrainian military mentions	0.461*** (0.102)				
Log civilian violence mentions		0.110*** (0.033)			
Log missile mentions			0.110*** (0.032)		
Log deescalation mentions				0.116*** (0.044)	
Log frontline mentions					0.449*** (0.091)
R2	0.610	0.601	0.601	0.598	0.613
N	675	675	675	675	675
Month FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

Table 3

Let's see.

```
~$ infile=Klymak_et_al_2025_Charitable_giving_in.pdf
```

```
~$ docker run -ti -v $(pwd)/$infile:/$infile -v  
$(pwd)/output:/output vpnagraj/tabulizer $infile 18
```

Table 3: Double machine learning for high-dimensional controls of donated amount

	(1)	(2)	(3)	(4)
	lasso-lasso	lasso-ridge	ridge-lasso	ridge-ridge
Log civilian casualties	0.14 (0.04) [0]	0.14 (0.04) [0]	0.13 (0.04) [0]	0.11 (0.03) [0]
Log military mentions	0.26 (0.05) [0.0]	0.18 (0.06) [0.0]	0.24 (0.05) [0.0]	0.14 (0.02) [0.02]

Table B.1

Last one.

```
$ infile=Klymak_et_al_2025_Charitable_giving_in.pdf
```

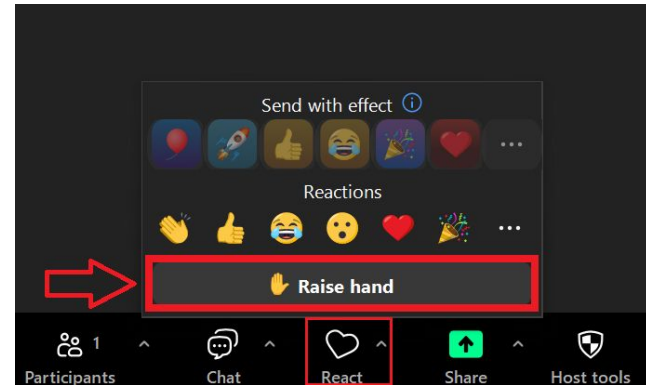
```
$ docker run -ti -v $(pwd)/$infile:/$infile -v  
$(pwd)/output:/output vpnagraj/tabulizer $infile 34
```

Table B.1: Summary statistics

	Mean	Standard deviation	Min	Max
Number of donations	3441	3489	281	54601
Log number of donations	7.98	0.49	5.64	10.91
Donated amount (UAH)	3,034,405	6,314,561	134,164	90,335,312
Log donated amount (UAH)	14.30	1.02	11.81	18.32
Total world events	116,089	34,831	40,518	212,164
Holidays	0.03	0.17	0.00	1.00
Come Back Alive events	0.07	0.25	0.00	1.00
Log Russian military casualties	2.22	0.88	-0.69	4.96
Log civilian casualties	3.12	0.65	0.69	5.38
Log air alert	2.54	1.44	-0.69	4.96
Log air strike	2.73	0.77	-0.69	5.17
Log art. shelling	4.36	0.62	2.56	6.04
Log hospital attack	0.20	0.91	-0.69	3.30
Log tank battles	0.73	1.22	-0.69	3.85
Log territory control claim	2.18	0.85	-0.69	4.55
Russia initiated event	109	75	18	526
Ukraine initiated event	43	41	6	260
Occupation	4	6	1	50
Log Ukrainian mentions	7.81	0.31	5.86	8.66
Log Ukrainian military mentions	6.20	0.33	3.87	7.21
Log civilian violence mentions	2.62	0.92	-0.69	5.43
Log missile mentions	2.61	0.93	-0.69	5.43
Log deescalation mentions	3.30	0.54	1.10	5.20
Log occupation mentions	2.17	0.70	-0.69	4.80
Log frontline mentions	5.93	0.35	3.50	7.00
Log sanctions	1.56	1.28	-0.69	4.06
Log total mentions	9.45	0.51	7.14	11.48
Log financial aid	15.58	2.06	14.90	27.53
Log humanitarian aid	13.44	3.08	11.81	24.17
Log military aid	13.56	3.02	12.37	26.15
Observations				

General instructions of the workshop

- “raise hands” if you have **completed the task**
- Questions in the chat
- Don't hesitate to ask!



Source: <https://tactiq.io/learn/how-to-raise-your-hand-in-zoom-meetings>

Hands-on training

- First: **tabulizer**
 - analyses/1_extraction_with_tabulizer_and_cleaning.qmd
- Second: **ocr (tesseract)**
 - analyses/2_extraction_with_tesseract_ocr_and_cleaning.qmd

Code will be written live for you to understand the process.

Schedule

- Presentation
- Install programs if necessary (docker + R packages) ~ 20 minutes
- Interval ~ 15 minutes
- Hands-on tabulizer
- Interval ~ 10 minutes
- Hands-on tesseract (ocr)
- Closing questions