

Scientific Data Curation: Efficiently extracting structured data from publications

Rupali Desai, Cytel, Pune, India

Nand Kishore Rawat, Cytel, Pune, India

ABSTRACT

In the current scenario of rapid pharmaceutical drug development, a large number of research papers are published every year. Research scientists find it increasingly challenging to keep pace with the deluge of information when they evaluate and compare technical notes between a compound being investigated and comparators such as an already approved drug, and when they compile data for meta-analysis. PubCoDe is a simple but highly effective software tool designed to assist scientists in extracting data from text, tables and plots in publications. PubCoDe supports intelligent text-pattern matching with keywords and contexts, assists in selective data picking from tables and performs rapid numeric data farming from graphs and figures. Extracted data are hyperlinked to the original source, thus providing a powerful tool to an independent validator. We have demonstrated that PubCoDe enhances the productivity of scientists in real-life scenarios.

INTRODUCTION

Research scientists in various scientific domains expend valuable time in the extraction of relevant data from published papers for various reasons. The volume of these relevant papers is vast and continuously increasing. This manual process of data extraction and validation is tedious and error-prone. PubCoDe is a software tool that addresses these challenges. It provides features such as searching publications for relevant terms, dragging text from publications and dropping it on to the outcomes database, extracting data from tables selectively along with applying transpose transformation, and extracting data from graphical images in publications. The PubCoDe software makes the process accurate and efficient. Further, PubCoDe enables validation of the extracted data by using hyperlinks located in the outcomes database. A click on the hyperlink in the outcomes database leads the user to the exact source location in the published papers. In order to validate the numbers extracted from graphs, the software provides features to reproduce the same graph with the extracted points and to perform a visual comparison, and to copy the extracted data and the reproduced graph to the required comparator document. This software is designed to read PDF (Portable Document Format) and Microsoft Word documents.

PROBLEM AND SOLUTION

The authors work with a Pharmacometrics team that receives assignments to extract relevant data from a given set of published papers.

The client typically provides a search protocol and the list of journals and published papers. The search protocol consists of search parameters such as patient characteristics including age, diagnosis, gender, treatment information, treatment duration, and inclusion and exclusion criteria. Next, research scientists (members of the Pharmacometrics team) use the PubCoDe software to search, extract and copy the relevant contents to the outcomes database. This database then can be used for further analysis of the data.

The scientific data extraction process encounters several challenges:

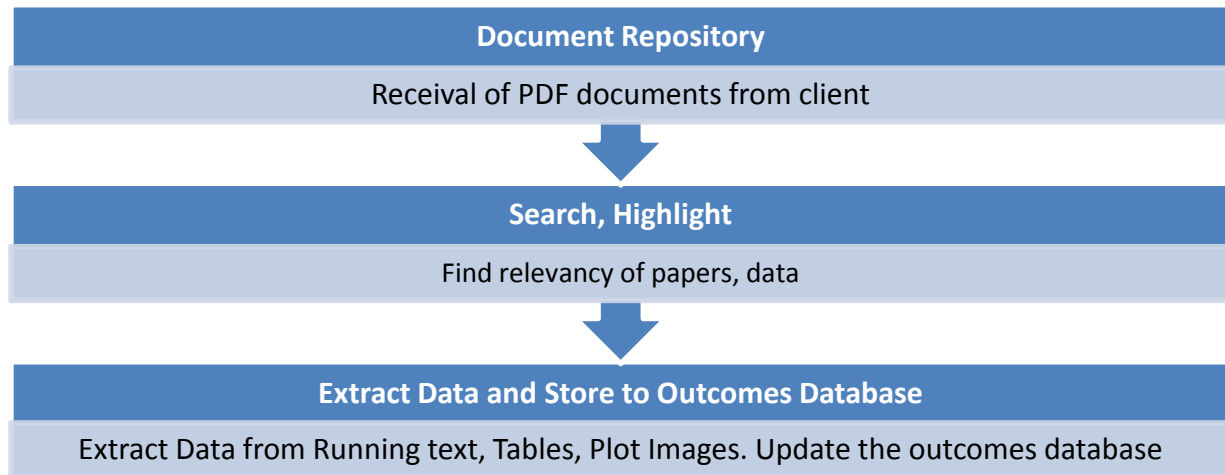
1. The manual process of finding relevant data is tedious.
2. The manual process can be dull and monotonous for research scientists.
3. The manual process is error-prone.
4. The manual process is slow, thus consuming the valuable time of research scientists.
5. The extraction of precise data points from a graphical image in the document is difficult.
6. The selective extraction of data from tables in the document is difficult.
7. The validation of the extracted data is tedious.

PhUSE 2016

PubCoDe overcomes these challenges by incorporating the following features:

1. Search with flexible options for case-sensitivity, whole words and word forms
2. Content-aware drag-and-drop from publications to outcomes database
3. Simultaneous search through all pages of the document
4. Multiple-term search in a single operation
5. Extraction of data points from a graph in the document
6. Easy data extraction from a table in the document
7. Availability of hyperlinks in the outcome database; hyperlinks lead to the exact location in the document

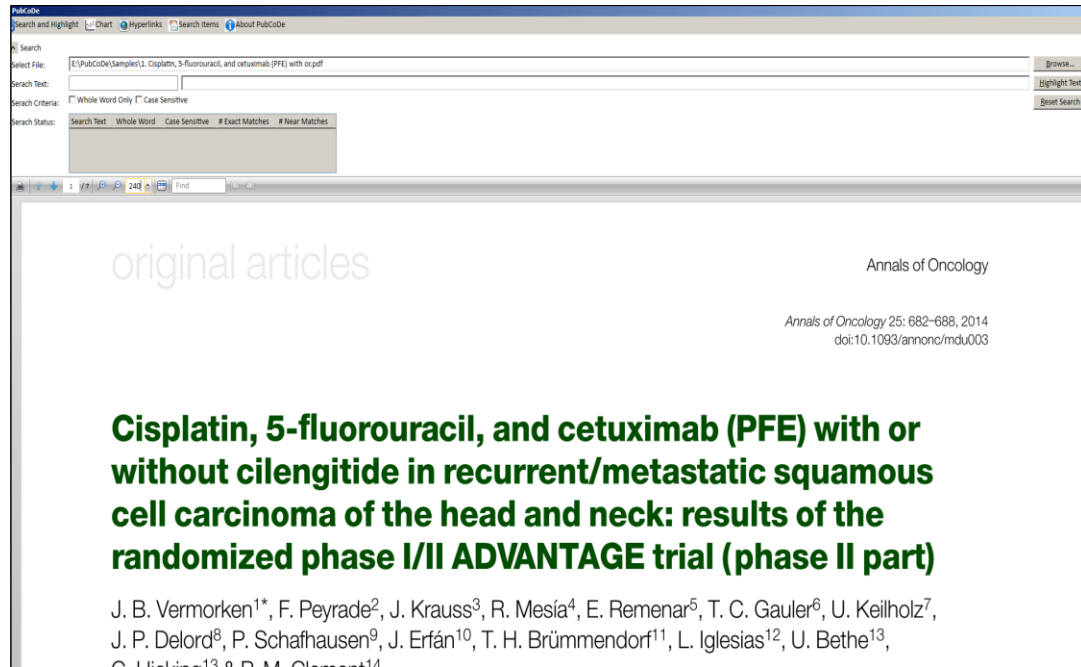
STEPS IN OPERATING PUBCODE



The features implemented in the software are as follows:

1. Open PDF/Word document:

The user can browse for the file containing the relevant data for the search.
After the document is opened, the user can perform basic document-viewer actions.



2. Highlight relevant single or multiple search terms:

- During the search and highlight operation, the user can provide multiple search terms. All the search terms are searched for and highlighted in the document in a single operation. This improves the efficiency by number of terms searched.
- All the pages in the document are scanned, and the search terms are highlighted to enable rapid reading.
- The number of matches for each word is listed in the output table in each row.
- Based on the search requirement, the user can select the option "Whole Word Only" and/or the option "Case Sensitive". The counts and the highlights will be updated accordingly in the output table.

Search

Select File: E:\PubCoDe\Samples\1. Cisplatin, 5-fluorouracil, and cetuximab (PFE) with or.pdf Browse...

Search Text: carcinoma|cetuximab Highlight Text

Search Criteria: ☐ Whole Word Only ☐ Case Sensitive Reset Search

Search Status:

Search Text	Whole Word	Case Sensitive	# Exact Matches	# Near Matches
carcinoma	☐	☐	11	0
cetuximab	☐	☐	27	2

Cisplatin, 5-fluorouracil, and cetuximab (PFE) with or without cilengitide in recurrent/metastatic squamous cell carcinoma of the head and neck: results of the randomized phase I/II ADVANTAGE trial (phase II part)

J. B. Vermorken^{1*}, F. Peyrade², J. Krauss³, R. Mesía⁴, E. Remenar⁵, T. C. Gauler⁶, U. Keilholz⁷, J. P. Delord⁸, P. Schafhausen⁹, J. Erfán¹⁰, T. H. Brümmendorf¹¹, L. Iglesias¹², U. Bethe¹³, C. Hicking¹³ & P. M. Clement¹⁴

3. Auto-complete for search terms:

- The user can store commonly searched terms in the internal repository once.
- Therefore, when the user types the letters in the search dialog, the software will provide a list of words starting with the typed letters as prompts.

PubCoDe

Search and Highlight | Chart | Hyperlinks | Search Items | About PubCoDe

Search

Select File: E:\PubCoDe\Samples\1. Cisplatin, 5-fluorouracil, and cetuximab (PFE) with or.pdf Browse...

Search Text: Treatment Highlight Text

Search Criteria: Treatment Reset Search

Search Status:

Search Text	Whole Word	Case Sensitive	# Exact Matches	# Near Matches
treat	☐	☐		
therapy	☐	☐		
therapies	☐	☐		
toxicity	☐	☐		
targets	☐	☐		
therapies	☐	☐		
Teriflunomide	☐	☐		

carcinoma

original articles

Annals of Oncology

4. Drag and drop text from document:

- The user can select text from the document and can start dragging the text out of the document to the outcome database.
- During the dragging action, the user can view the text being dragged. This feature facilitates accurate data extraction.

treatment

KPS (<80 versus ≥ 80)

Patients were stratified according to their KPS (<80 versus ≥80) and randomized 1:1:1 to 1 of the 3 parallel treatment arms: PFE + CIL1W, PFE + CIL2W, and PFE (Figure 1). See supplementary Data, available at

Recurrent and/or metastatic squamous cell carcinoma of the head and neck (R/M-SCCHN) over-expresses $\alpha v \beta 5$ integrin. Cilengitide selectively inhibits $\alpha v \beta 3$ and $\alpha v \beta 5$ integrins and is investigated as a treatment strategy.

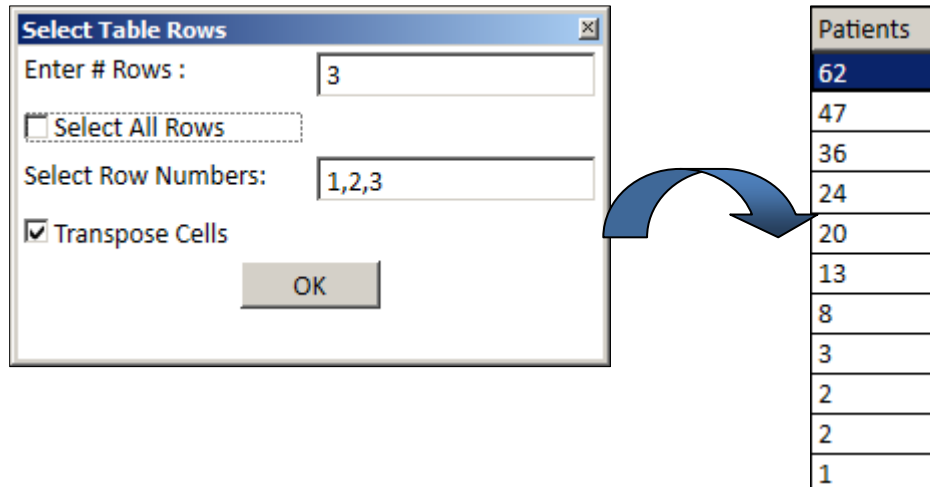
Background: Recurrent and/or metastatic squamous cell carcinoma of the head and neck (R/M-SCCHN) over-expresses $\alpha v \beta 5$ integrin. Cilengitide selectively inhibits $\alpha v \beta 3$ and $\alpha v \beta 5$ integrins and is investigated as a treatment strategy.

5. Extract and parse information from tables:

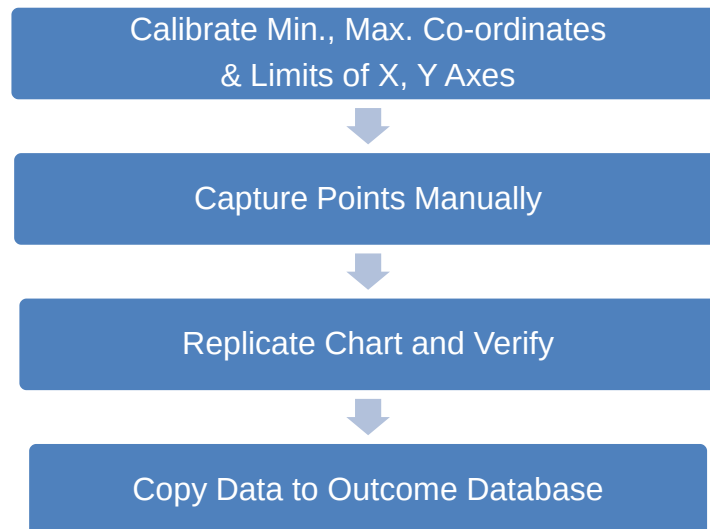
- The user can select tabular data from the document and can then click on the table icon  on the toolbar.
- The user must enter the number of rows selected.
- The user will specify the rows to be copied and will indicate whether the rows and columns must be transposed. All the rows or only selected rows can be copied.
- The selected rows are copied to the clipboard. Then, they can be pasted into the outcomes database.

Number of patients at risk

PFE	62	47	36	24	20	13	8	3	2	2	1
PFE + CIL1W	62	51	42	25	16	9	6	2	0	0	0
PFE + CIL2W	60	42	32	13	7	4	2	0	0	0	0



6. Read from graphs and create numeric data columns:



The graph image from the PDF document can be saved as an image file. This software feature works with the image file.

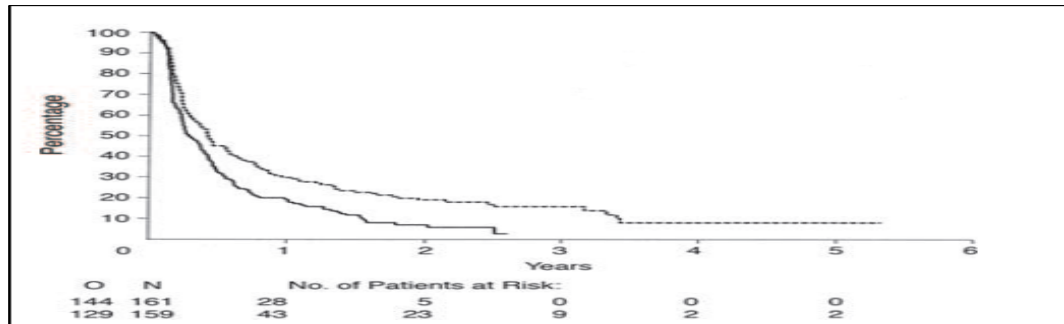
- After the image is imported into the software, the first action required is the calibration of the minimum and maximum points of the X and Y axes. The user calibrates the X and Y axes.
- The user specifies the limits for both the axes.
- The user hovers the mouse, checks the values and selects the points by using mouse clicks. Every mouse click will add a data point to the table.
- The user regenerates the plot image by using the captured points from the software.
- The user copies the data from the table to the outcomes database.
- This feature ensures accurate extraction of data points from graph images.

The steps are described in detail below:

i. Calibrate X,Y Axes:

PhUSE 2016

- Calibrate XMin, XMax, YMin and YMax co-ordinates by selecting the point on the screen.
- The point is selected accurately by referring to the zoomed view.



Reset Table
Reset Chart
Plot Points
Copy Chart

Select Image File: E:\PubCoDe\Samples\Figure 1.bmp Browse...

Calibrate Axes:

XMin	XMax	YMin	YMax	Reset
80, 281.82	485, 282.82	80, 282.82	82, 27.82	

Reference X, Y Screen Coordinates:

Enter X, Y Axes Limits:

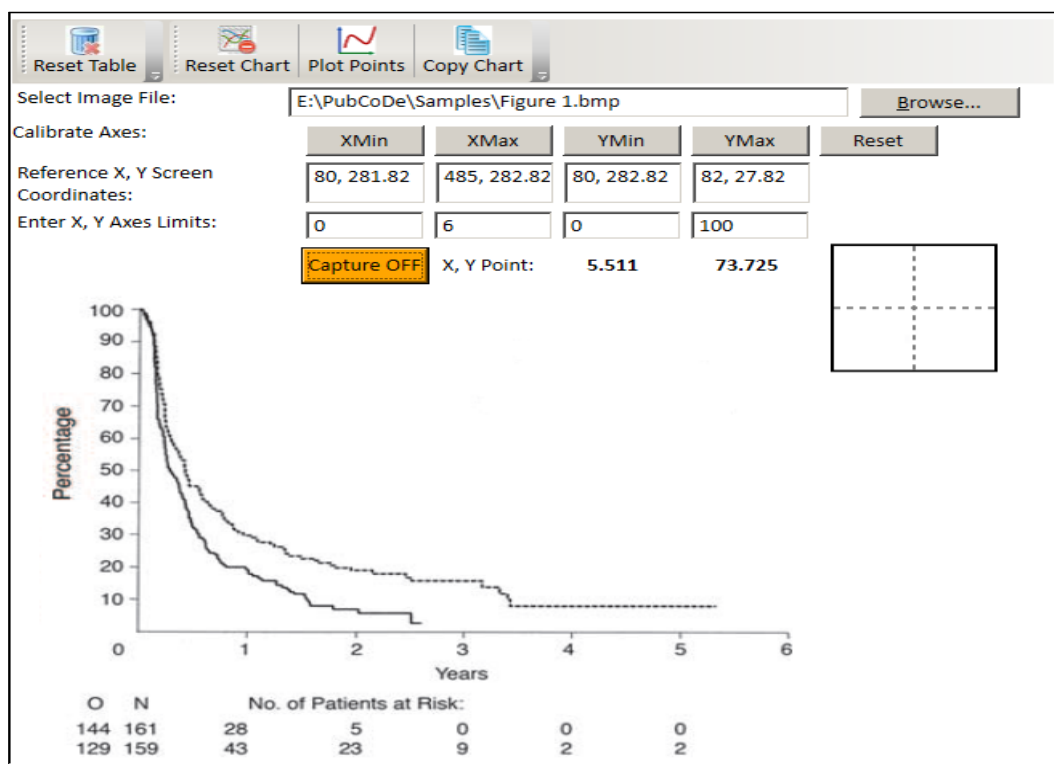
0	0	0	0
---	---	---	---

Capture ON Co-ordinates 104 129.82

ii. Enter Limits of X, Y Axes:

- Enter the minimum and maximum values of the X and Y axes in the Axes Limits edit boxes.

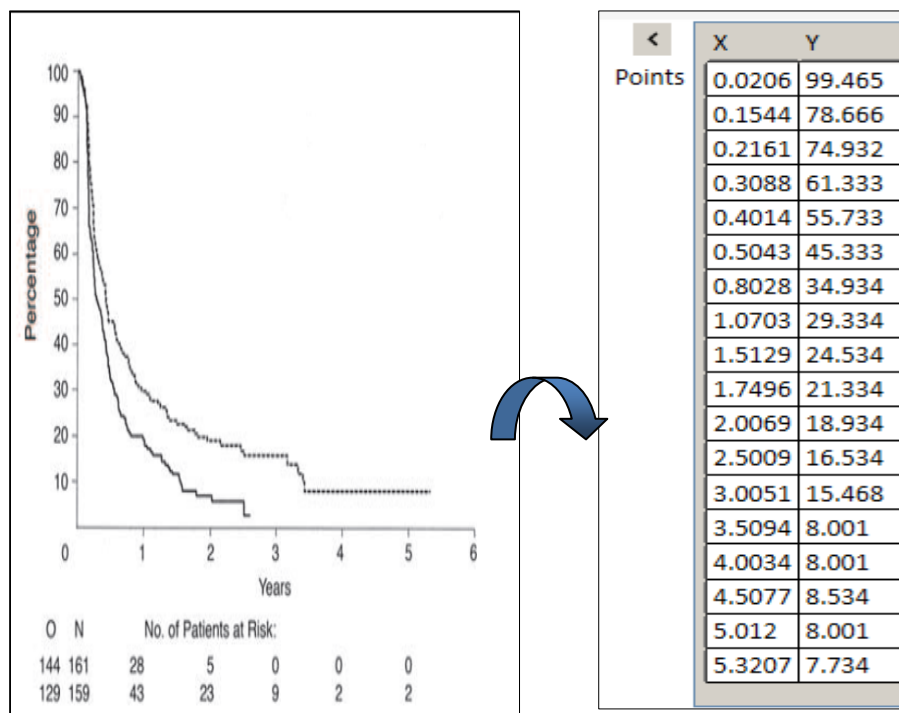
PhUSE 2016



iii. Capture Points:

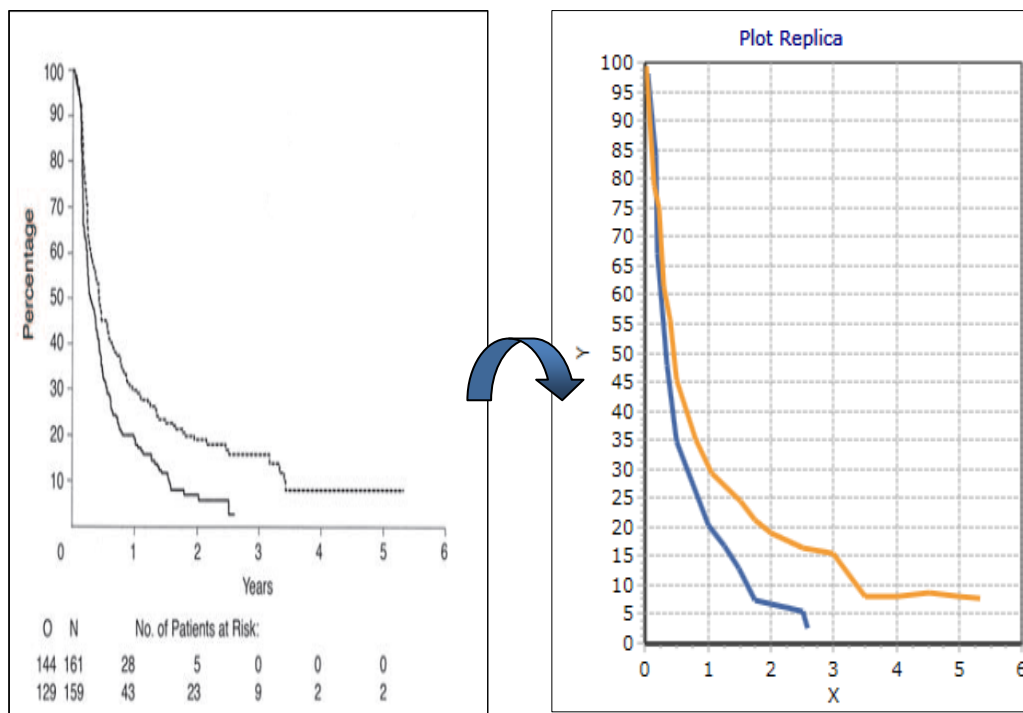
- Click the "Capture ON" button. Hover the mouse over the image; the X, Y values will be displayed in the labels before the box.
- The selection can be performed accurately by using the zoomed image in the box.
- After the required X, Y co-ordinates are obtained, click the mouse. This operation will result in point selection. These points are registered in the table.

PhUSE 2016



iv. Replicate Plot and Verify:

- Click the "Plot Points" button to reproduce the plot.
- Verify whether the reproduced plot is the same as the original plot image.



v. Copy the Data to the Outcomes Database:

PhUSE 2016

- Select the X, Y columns from the table, perform a right-click and select “Copy”.
- The user can paste these data points into the outcomes database.

vi. Copy Chart:

- Copy the reproduced plot to the clipboard.
- This copied plot can be pasted into the outcomes database.

vii. Reset Chart:

- Click “Reset Chart” to reset the chart to a blank one.

viii. Reset Table:

- Click “Reset Table” to clear the captured points in order to restart the capturing of points.

7. Validate extracted data using links:

After the data are extracted by the data curation team, the extracted data can be validated by using links present within the software. This feature allows the independent validator to verify the accuracy of data.

Reference	Phase	Treatment Arm	patients
Cisplatin, 5-fluorouracil, and cetuximab (PFE)	phase I/II	PFE + CIL1W	62
	phase I/II	PFE + CIL1W	51
	phase I/II	PFE + CIL1W	42
	phase I/II	PFE + CIL1W	25
	phase I/II	PFE + CIL1W	16
	phase I/II	PFE + CIL1W	9
	phase I/II	PFE + CIL1W	6
	phase I/II	PFE + CIL1W	2
	phase I/II	PFE + CIL1W	0
	phase I/II	PFE + CIL1W	0
	phase I/II	PFE + CIL1W	0

Cisplatin, 5-fluorouracil, and cetuximab (PFE) with or without cilengitide in recurrent/metastatic squamous cell carcinoma of the head and neck: results of the randomized phase I/II ADVANTAGE trial (phase II part)

J. B. Vermorken^{1*}, F. Peyrade², J. Krauss³, R. Mesía⁴, E. Remenar⁵, T. C. Gauler⁶, U. Keilholz⁷, J. P. Delord⁸, P. Schaffhausen⁹, J. Erfán¹⁰, T. H. Brümmendorf¹¹, L. Iglesias¹², U. Bethe¹³, C. Hicking¹³ & P. M. Clement¹⁴

Reference	Phase	Treatment Arm	patients
Cisplatin, 5-fluorouracil, and cetuximab (PFE)	phase I/II	PFE + CIL1W	62
	phase I/II	PFE + CIL1W	51
	phase I/II	PFE + CIL1W	42
	phase I/II	PFE + CIL1W	25
	phase I/II	PFE + CIL1W	16
	phase I/II	PFE + CIL1W	9
	phase I/II	PFE + CIL1W	6
	phase I/II	PFE + CIL1W	2
	phase I/II	PFE + CIL1W	0
	phase I/II	PFE + CIL1W	0
	phase I/II	PFE + CIL1W	0



Months											
Number of patients at risk											
PFE	62	47	36	24	20	13	8	3	2	2	1
PFE + CIL1W	62	51	42	25	16	9	6	2	0	0	0
PFE + CIL2W	60	42	32	13	7	4	2	0	0	0	0

CONCLUSION

The principle of PubCoDe software is extremely simple; the software allows quick extraction of data from electronic publications in an accurate and contextually intelligent manner. The application of this software in a real-life scenario demonstrates that PubCoDe can save a significant amount of time for researchers, thus greatly improving the efficiency and quality of scientific data curation in pharmaceutical drug development and other scientific domains. The current version of the software can read PDF and Microsoft Word documents; in future work, the authors will consider implementing an enhanced version that can process image documents by using Optical Character Recognition (OCR).