The Royal College of Pathologists
Pathology: the science behind the cure

Microsoft Excel as a data extraction tool for audit of cancer minimum datasets: a brief how-to guide

This brief guide accompanies an article published in the Bulletin in January 2019 entitled *Microsoft Excel as a data extraction tool for audit of cancer minimum datasets*. As the article states, the histopathology archive of cancer biopsy reports is a potentially rich source of data for audit and research; however, much of this is locked within pathology IT systems.

Read in conjunction with its corresponding *Bulletin* article, this guide aims to provide you with a little more understanding of how extract the text of standardised breast cancer reports into a spreadsheet and use the sophisticated formulae available in Microsoft Excel to audit cancer minimum datasets.

The example here relates to extracting ER and PR positivity rates for breast cancer.

Extracting the biopsy reports

The rather antiquated pathology IT systems of many departments can perform searches by SNOMed code and any other coded item. Using this search function, it is possible to extract all breast cancer core biopsies including the text of the reports over two-month intervals. Instead of printing to a printer the system prints to the 'host file system' ('HFS') which retains an electronic version of the whole search. This search can then be opened in other formats, such as Excel.

Patient identifiers such as name or date of birth should be avoided in the initial data extraction. The specimen accession number is usually the most appropriate identifier to use in the download of large amounts of data.

It is fundamental to this approach that the original reports are in a standardised format used by the whole team of reporting pathologists. Standardised reports with each data item on a new line enable searches to be carried out for specific data.

When a report is downloaded into Excel, each line forms a new row in the spreadsheet. However, if the report is in a free text style, individual data items may straddle two lines or several items may appear in the same line of text, rendering retrieval more problematic. Reporting proformas can be useful to allow for free text when this is essential, but as an addition to the proforma data – that way it will not impede subsequent extraction of the key items.

Using formulae in Excel to search the reports

Although constructing the formulae within Excel may seem daunting, there is readily accessible guidance on their design on the internet. There may be several standard Excel formulae that can be adapted depending on the data item being searched for and its format within the report. Any formulae will need to be tailored for the item being searched for and its context.

After extracted, data should be sense-checked due to the risk of unintended data being included in the results. For example, a basic search for ER results using the occurrence of the letters 'ER'

will result in these letters being identified within common words. By building in other search criteria, the data can be honed to precisely the required item.

In the example in Table 1, all the text of the reports is in column A and formulae are inserted into the subsequent columns to search the contents of the text report. The formula in cell B1 searches the text in A1 for a match to the specified text ‘H,18.00’ and allocates a ‘1’ if the match is present. Use of the conditional formatting tools in Excel will highlight each new laboratory number and associated report. Counting all the cells containing ‘1’ in column B (using =COUNTIF(B1:B5, 1) in a separate part of the spreadsheet calculates the total number of cases reviewed.

Table 1: Example of formulae used to search for report data

	B1	*fx*	=SUMPRODUCT(--ISNUMBER(SEARCH({"H,18.00"},A1)))						
	A		B	C	D	E	F	G	H
1	H,18.00999999		1	0					
2	ER positive, Quickscore (5/5 + 3/3 = 8/8)		0	3					
3	Positive with cytokeratin stains.		0	2					
4	Her2 requested		0	1					
5			0	0					

In Table 2, the formulae in column C search the cells in column A for three data items and allocate a ‘1’ to each when present. Notice that the formula recognises the word ‘positive’ and the letters ‘er’ in ‘cytokeratin’ in the free text in cell A3 and ‘Her2’ in A4, but, in the absence of the ‘8/8’ the full score of 3 is not achieved. Counting the occasions that the number 3 occurs in column C – by using =COUNTIF(C1:C5,3) – will give the total number of cases that are ER positive with a score of 8/8.

Table 2: Example of formulae used to search for multiple report data

	C3	*fx*	=SUMPRODUCT(--ISNUMBER(SEARCH({"ER","positive","8/8"},A3)))						
	A		B	C	D	E	F	G	H
1	H,18. 00999999		1	0					
2	ER positive, Quickscore (5/5 + 3/3 = 8/8)		0	3					
3	Positive with cytokeratin stains.		0	2					
4	Her2 requested		0	1					

Within Excel there is an option to download all reports into one column of the spreadsheet or convert the report into multiple columns with each word of the report in a separate column – this process uses the space in the text as a ‘delimiter’. This function is found in the ‘Data’ and ‘Text to columns’ tabs. In Table 3, the pathology report has been separated into multiple columns using this process.

If a standardised proforma report has been used, the same data item will appear consistently in the same column, greatly enhancing its extraction from the spreadsheet. The formula in cell M6 relies on the consistent appearance of the text ‘grade:’ in column B to extract the number in column C. This allows large-scale analysis of the grading of breast cancer.

Table 3: Example of words in report split into separate columns

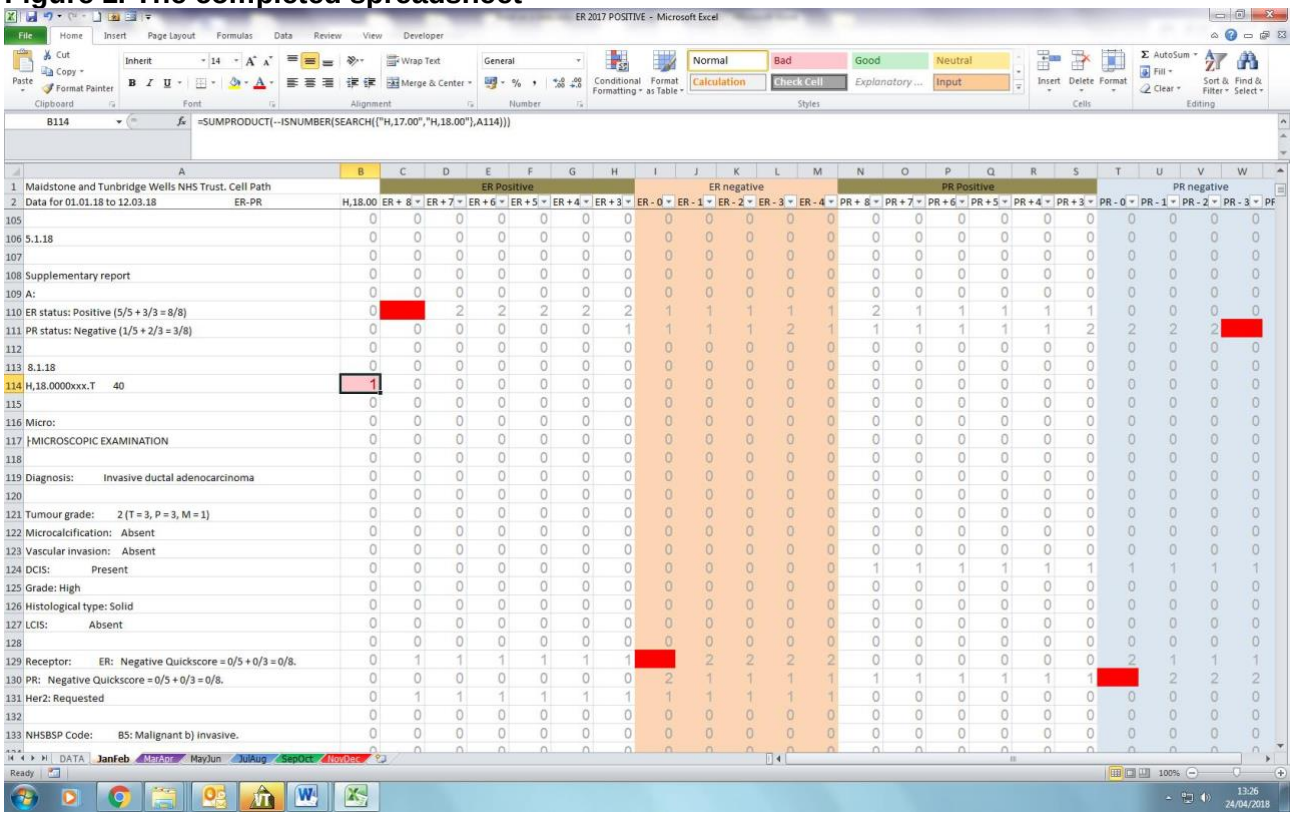
	M6	*fx*	=IF(B6="grade:",C6,"…")										
	A	B	C	D	E	F	G	H	I	J	K	L	M
5	Diagnosis:	Invasive	ductal	carcinoma	NST								"…"
6	Tumour	grade:	3	(T	=	3,	P	=	3,	M	=	3)	3

Formulae can be copied into adjacent columns and rows and then adapted to ensure that the correct item is searched for. The completed ER and PR spreadsheet is seen in Figure 1 below. Conditional formatting is used to highlight data of interest.

The red flags indicate ER and PR results, where a score of 3 was recorded for the text in the adjacent 'A' cell. The pink flag highlights the start of the next case by identifying the specimen number in column A.

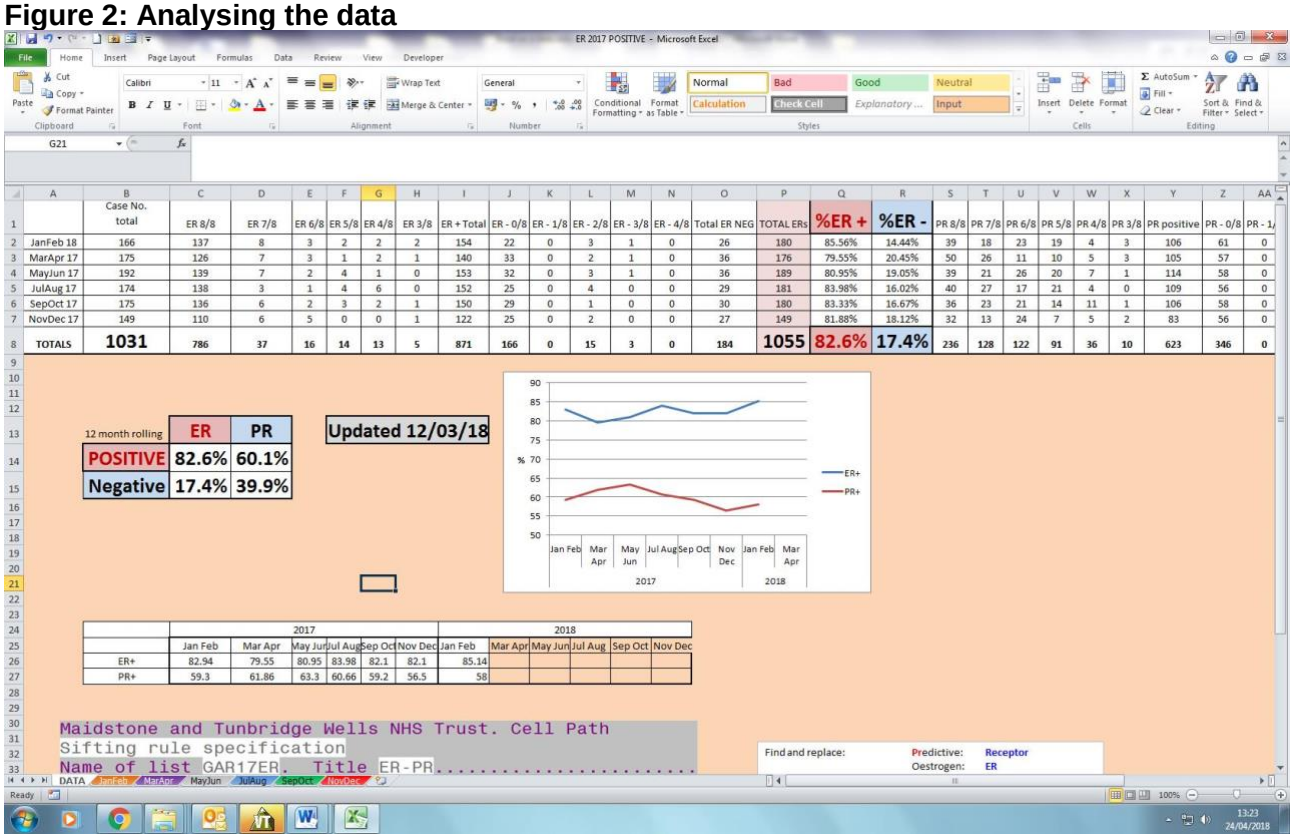
The coloured flags aid sense-checking of the data by allowing the user to rapidly scan the sheet to make sure each cancer case has an ER and a PR result.

Figure 1. The completed spreadsheet



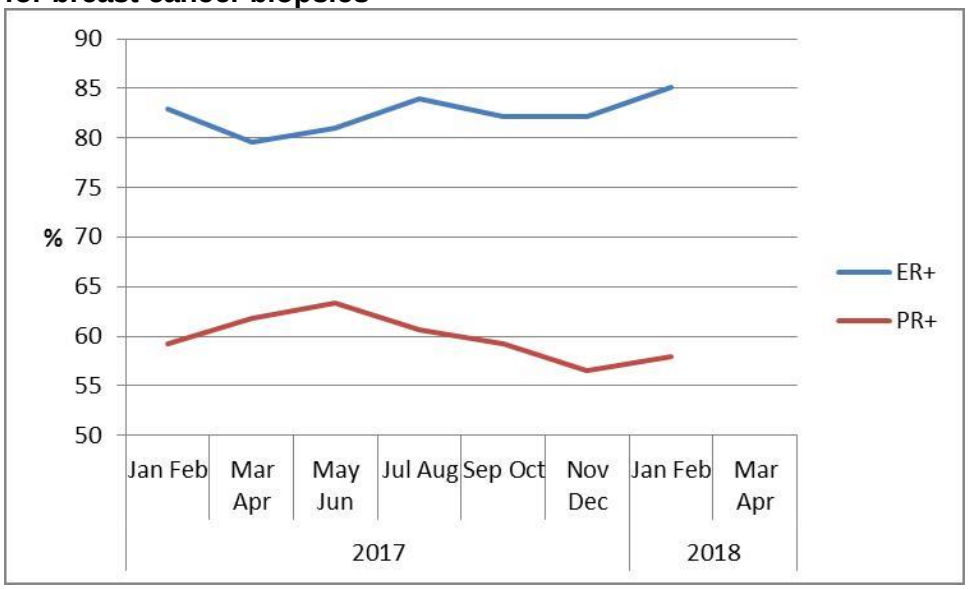
A separate spreadsheet can be used to analyse the data in the columns (see Figure 2). This is a separate datasheet within the ER/PR spreadsheet used to analyse the data extracted from the breast cancer reports. Including data of individual ER/PR scores allows the pattern and intensity of staining to be scrutinised, in the case that there is an issue with the quality of the staining.

Figure 2: Analysing the data



This approach to data analysis has the added advantage that it can be saved and updated to permit a rolling 12-month audit of the data items. Figure 3 shows the results of the audit plotted on a line graph. This data is updated and added to the sheet every two months.

Figure 3: Results of a 12-month rolling audit of ER and PR positivity rates for breast cancer biopsies



This brief how-to guide was written by Dr Graham Russell, Consultant Histopathologist, Maidstone and Tunbridge Wells NHS Trust.