Egon Willighagen is one of the of people whose support has kept me going through tough times. Here he supports my criticism of PDF (No, PDFs really do suck!). For background: I had posted a criticism of PDF and a lively discussion took place on FriendFeed. I understand it is questionable as to whether FF discussions should be quoted – they are public but can be taken out of context – so I didn't. I had a few supporters, but a surprising number of naysayers, the general tenor of which was that PDF is easy to read and HTML is difficult to read. So I've done an experiment, but first Egon's post:

No, PDFs really do suck!

EW: A typical blog by Peter MR made (again), The ICE-man: Scholary HTML not PDF, the point of why PDF is to data what a hamburger is to a cow, in reply to a blog by Peter SF, Scholarly HTML.

This lead to a discussion on FriendFeed. A couple of misconceptions:

FF: ***"But how are we going to cite without paaaaaaaaaaage nuuuuuuuuuumbers?"***

EW: We don't. Many online-only journals can do without; there is DOI. And if that is not enough, the legal business has means of identifying paragraphs, etc, which should provide us with all the methods we could possibly need in science.

FF: ***Typesetting of PDFs, in most journals, is superior than HTML, which is why I prefer to read a PDF version if it is available. It is nicer to the eyes.***

EW:Ummm... this is supposed to be Science, not a California Glossy. It seems that pretty looks is causing major body count in the States. Otherwise, HTML+CSS can likely beat any pretty looks of PDF, or at least match it.

*FF:****As I seem to be the only physicist/mathematician who comments on these sort of things, I feel like a broken record, but math support in browsers currently sucks extremely badly and this is a primary reason why we will continue to use PDF for quite some time.***

EW: HTML+MathML is well established, and default FireFox browsers have no problem showing mathematical equations. For years, the Blue Obelisk QSAR descriptor ontology has been using such a set up for years. If you use TeX to author your equations, you can convert it to HTML too.

*FF:****We can mine the data from the PDF text.***

EW:Theoretically, yes. Practically, it is money down the drain. PDF is particularly nasty here, as it breaks words at the end of a line, and even can make words consist of unlinked series of characters positioned at (x,y). PDF, however, can contains a lot of metadata, but that is merely a hack, and unneeded workaround. Worse, hardly used regarding chemistry. PDF can contain PNG images which can contain CML; the tools are there, but not used, and there are more efficient technologies anyway.

EW: I, for one, agree with Peter on PDF: it really suck as scientific communication medium.

So here's an experiment with a sample size of one. I went to BiomedCentral, took the first journal I came across which had a ta-ble. (A ta-ble is a coll-ect-ion of num-bers in rows and col-umns). Sometimes I read tables, but sometimes I put them into a spread-sheet. (A spread-sheet is soft-ware that lets you cal-cul-ate things). The article is http://www.biomedcentral.com/1471-2105/9/545 which is called the full-text and is in HTML. I went to Table 1 and found:

Table 1

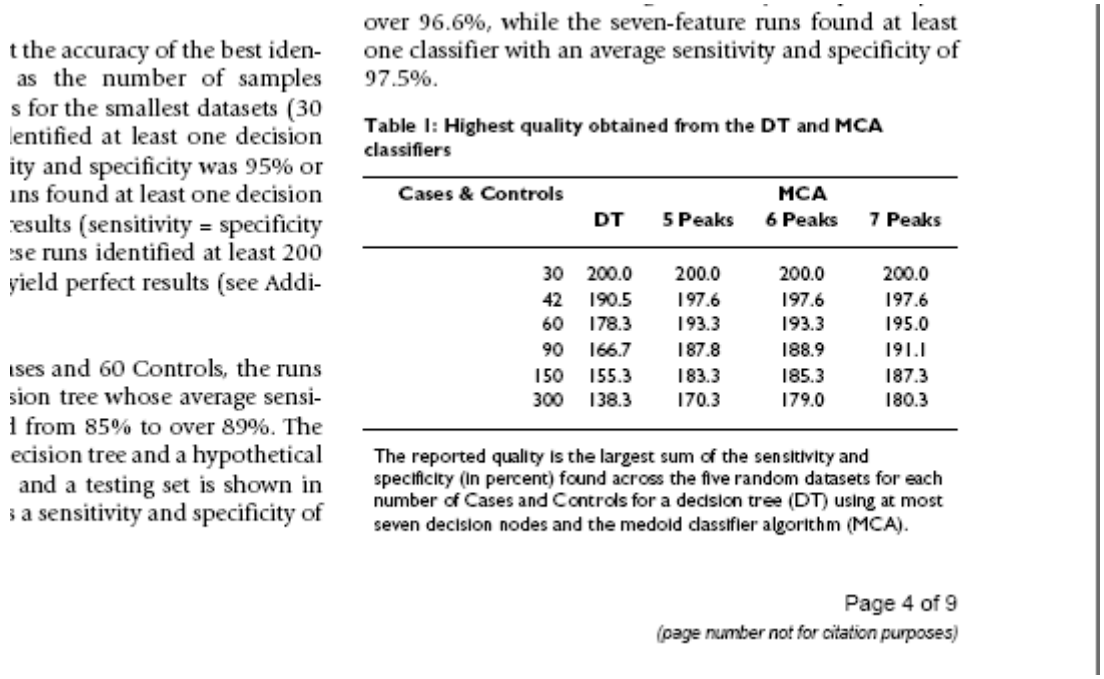
Highest quality obtained from the DT and MCA classifiers

Cases & Controls	DT	MCA 5 Peaks	MCA 6 Peaks	MCA 7 Peaks
30	200.0	200.0	200.0	200.0
42	190.5	197.6	197.6	197.6
60	178.3	193.3	193.3	195.0
90	166.7	187.8	188.9	191.1
150	155.3	183.3	185.3	187.3
300	138.3	170.3	179.0	180.3

The reported quality is the largest sum of the sensitivity and specificity (in percent) found across the five random datasets for each number of Cases and Controls for a decision tree (DT) using at most seven decision nodes and the medoid classifier algorithm (MCA).

Luke and Collins *BMC Bioinformatics* 2008 **9**:545 doi:10.1186/1471-2105-9-545

I then went to the PDF (which is not seen by BMC as the full-text) http://www.biomedcentral.com/content/pdf/1471-2105-9-545.pdf and found the same table:



t the accuracy of the best iden-
as the number of samples
s for the smallest datasets (30
lentified at least one decision
ity and specificity was 95% or
ıns found at least one decision
esults (sensitivity = specificity
se runs identified at least 200
yield perfect results (see Addi-

ıses and 60 Controls, the runs
sion tree whose average sensi-
l from 85% to over 89%. The
ecision tree and a hypothetical
and a testing set is shown in
s a sensitivity and specificity of

over 96.6%, while the seven-feature runs found at least one classifier with an average sensitivity and specificity of 97.5%.

Table 1: Highest quality obtained from the DT and MCA classifiers

Cases & Controls	DT	MCA 5 Peaks	MCA 6 Peaks	MCA 7 Peaks
30	200.0	200.0	200.0	200.0
42	190.5	197.6	197.6	197.6
60	178.3	193.3	193.3	195.0
90	166.7	187.8	188.9	191.1
150	155.3	183.3	185.3	187.3
300	138.3	170.3	179.0	180.3

The reported quality is the largest sum of the sensitivity and specificity (in percent) found across the five random datasets for each number of Cases and Controls for a decision tree (DT) using at most seven decision nodes and the medoid classifier algorithm (MCA).

Page 4 of 9
(page number not for citation purposes)

[I have reproduced them at the same size as they cam up in my open-source browser. The HTML was rendered naturally by the browser with no help from me. The PDF required me to download a closed-source proprietary plugin from Adobe.

I am not an expert on readability but I would like to see the researched arguments that says the HTML is worse for humans than the PDF (actually I think it's better).

But here is the clincher. As a scientist I don't just want to read the paper with my eyes. I want to use the numbers. Maybe I want to see how column 1 varies with column 2. The natural way to do this is to cut-and-paste the table. (This is done in each case by sweeping out the table with the cur-sor and pressing the Ctrl-key and the C key on the key-board and the same time. The data is now on the clip-board). I then open up Ex-cel (because I am in the pay of Mi-cro-soft) and "paste" the clipboard into the spread-sheet. This is what I get from the PDF version.

Table 1: Highest quality obtained from the DT and MCA
classifiers
Cases & Controls MCA
DT 5 Peaks 6 Peaks 7 Peaks
30 200.0 200.0 200.0 200.0
42 190.5 197.6 197.6 197.6
60 178.3 193.3 193.3 195.0
90 166.7 187.8 188.9 191.1
150 155.3 183.3 185.3 187.3
300 138.3 170.3 179.0 180.3
The reported quality is the largest sum of the sensitivity and
specificity (in percent) found across the five random datasets for each
number of Cases and Controls for a decision tree (DT) using at most
seven decision nodes and the medoid classifier algorithm (MCA).

All the data has gone into one column. The tabular nature has been completely destroyed. And the cut-and-paste was done with Adobe's own tool, so even Adobe doesn't know what a table is in the PDF. (I have been taken to task for criticizing PDF because some people don't use Adobe tools).

Here's the HTML version. I have highlighted a cell to show that all cells are in correct columns:

Table 1

Highest quality obtained from the DT and MCA classifiers

Cases & Controls		MCA		
	DT	5 Peaks	6 Peaks	7 Peaks
30	200	200	200	200
42	190.5	197.6	197.6	197.6
60	178.3	193.3	193.3	195
90	166.7	187.8	188.9	191.1
150	155.3	183.3	185.3	187.3
300	138.3	170.3	179	180.3

The reported quality is the largest sum of the sensitivity and specificity (in percent) found across the five random datasets for each number of Cases and Controls for a decision tree (DT) using at most seven decision nodes and the medoid classifier algorithm (MCA).

Luke and Collins *BMC Bioinformatics* 2008 **9**:545
doi:10.1186/1471-2105-9-545

I was called “a bit … dogmatic”. Yes I am. This seems to me so self-evidently a case for using HTML over PDF that I can't think of any reason why PDF should be used.

And kudos to BMC. They have realised that HTML is a better digital medium than PDF. Are their readers cancelling subscriptions? No...

Oops... BMC is an Open Access publisher. It is forcing its authors to pay for their manuscripts to be converted into horrid HTML. I expect they'll start sending their papers elsewhere...

Xxx