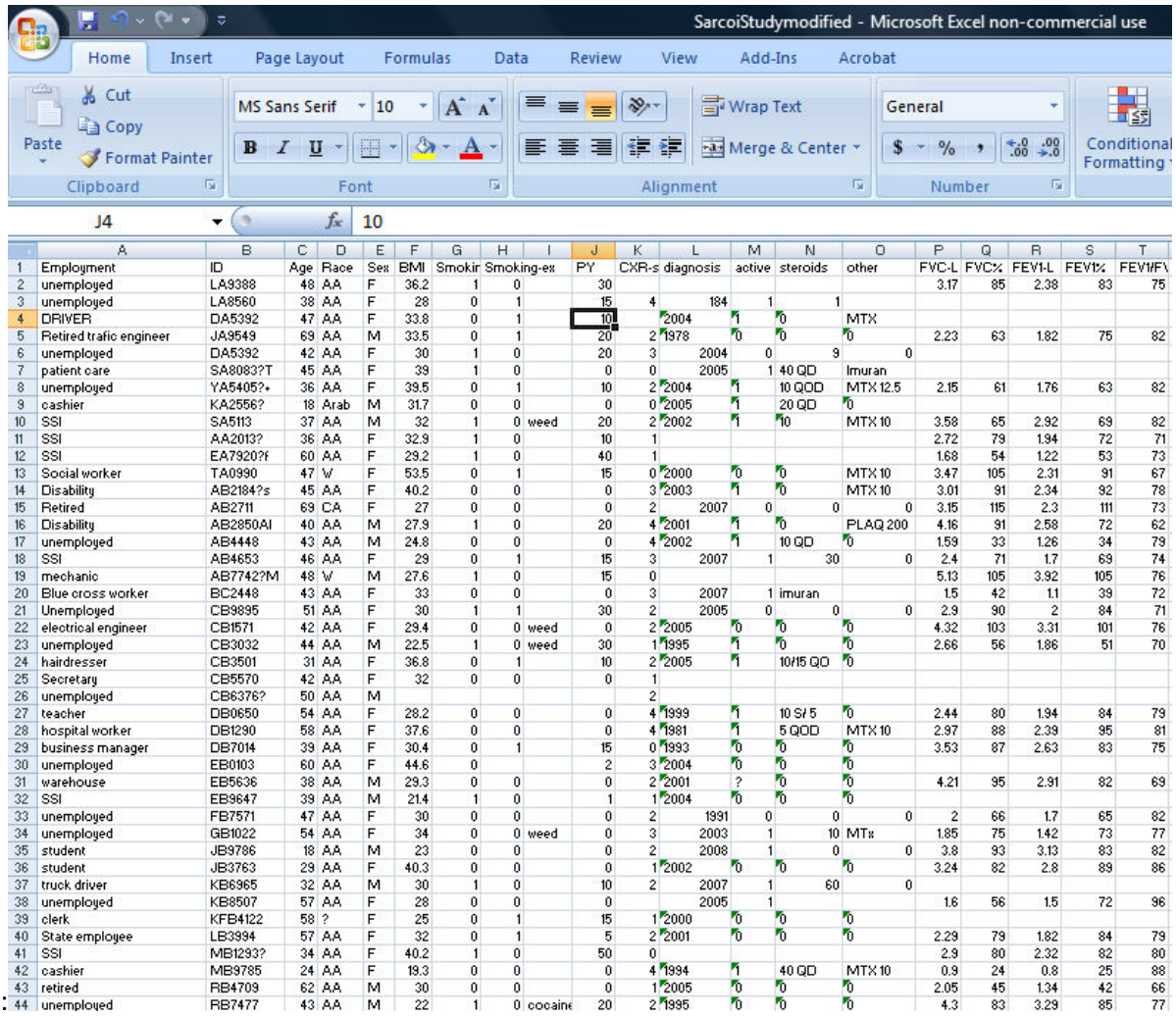
Using JMP® and SAS® Software to Prepare a Last-Minute Research Project for Publication
or
Statistical Modeling of Breathing Relief in Sarcoidosis Patients

Jim Woods, Wayne State University, Detroit, MI

A medical school student presented her data in Excel to a research assistant, and asked that the research assistant quickly process the data to provide some results that could be submitted as a draft to a magazine, for publication. Doctors *do* like to get published at universities, and there was a deadline, and it's due today! The lowly research assistant consoled the anxious doctor that there yet be hope, because "SAS® Saves Time" and JMP® is "Statistical Discovery Software," and JMP® can work with SAS®, so problem solved! Can the student make the publication deadline?

The data was provided in Excel as follows, and immediately the assistant saw problems—some missing data, some character/numeric mixups, some missing labels, some vague descriptions,



	Employment	ID	Age	Race	Sex	BMI	Smokir	Smoking-ex		PY	CXR-s	diagnosis	active	steroids	other	FVC-L	FVC%	FEV1-L	FEV1%	FEV1/F
2	unemployed	LA9388	48	AA	F	36.2	1	0		30						3.17	85	2.38	83	75
3	unemployed	LA8560	38	AA	F	28	0	1		15	4	184	1	1						
4	DRIVER	DA5392	47	AA	F	33.8	0	1		10		2004	1	0	MTX					
5	Retired trafic engineer	JA9549	69	AA	M	33.5	0	1		20	2	1978	0	0	0	2.23	63	1.82	75	82
6	unemployed	DA5392	42	AA	F	30	1	0		20	3	2004	0	9	0					
7	patient care	SA8083?T	45	AA	F	39	1	0		0	0	2005	1	40 QD	Imuran					
8	unemployed	YA5405?•	36	AA	F	39.5	0	1		10	2	2004	1	10 QOD	MTX 12.5	2.15	61	1.76	63	82
9	cashier	KA2556?	18	Arab	M	31.7	0	0		0	0	2005	1	20 QD	0					
10	SSI	SA5113	37	AA	M	32	1	0	weed	20	2	2002	1	10	MTX 10	3.58	65	2.92	69	82
11	SSI	AA2013?	36	AA	F	32.9	1	0		10	1					2.72	79	1.94	72	71
12	SSI	EA7920?f	60	AA	F	29.2	1	0		40	1					1.68	54	1.22	53	73
13	Social worker	TA0990	47	W	F	53.5	0	1		15	0	2000	0	0	MTX 10	3.47	105	2.31	91	67
14	Disability	AB2184?s	45	AA	F	40.2	0	0		0	3	2003	1	0	MTX 10	3.01	91	2.34	92	78
15	Retired	AB2711	69	CA	F	27	0	0		0	2	2007	0	0	0	3.15	115	2.3	111	73
16	Disability	AB2850AI	40	AA	M	27.9	1	0		20	4	2001	1	0	PLAQ 200	4.16	91	2.58	72	62
17	unemployed	AB4448	43	AA	M	24.8	0	0		0	4	2002	1	10 QD	0	1.59	33	1.26	34	79
18	SSI	AB4653	46	AA	F	29	0	1		15	3	2007	1	30	0	2.4	71	1.7	69	74
19	mechanic	AB7742?M	48	W	M	27.6	1	0		15	0					5.13	105	3.92	105	76
20	Blue cross worker	BC2448	43	AA	F	33	0	0		0	3	2007	1	imuran		1.5	42	1.1	39	72
21	Unemployed	CB9895	51	AA	F	30	1	1		30	2	2005	0	0	0	2.9	90	2	84	71
22	electrical engineer	CB1571	42	AA	F	29.4	0	0	weed	0	2	2005	0	0	0	4.32	103	3.31	101	76
23	unemployed	CB3032	44	AA	M	22.5	1	0	weed	30	1	1995	1	0	0	2.66	56	1.86	51	70
24	hairdresser	CB3501	31	AA	F	36.8	0	1		10	2	2005	1	10/15 QD	0					
25	Secretary	CB5570	42	AA	F	32	0	0		0	1									
26	unemployed	CB6376?	50	AA	M						2									
27	teacher	DB0650	54	AA	F	28.2	0	0		0	4	1999	1	10 S/ 5	0	2.44	80	1.94	84	79
28	hospital worker	DB1290	58	AA	F	37.6	0	0		0	4	1981	1	5 QOD	MTX 10	2.97	88	2.39	95	81
29	business manager	DB7014	39	AA	F	30.4	0	1		15	0	1993	0	0	0	3.53	87	2.63	83	75
30	unemployed	EB0103	60	AA	F	44.6	0			2	3	2004	0	0	0					
31	warehouse	EB5636	38	AA	M	29.3	0	0		0	2	2001	?	0	0	4.21	95	2.91	82	69
32	SSI	EB9647	39	AA	M	21.4	1	0		1	1	2004	0	0	0					
33	unemployed	FB7571	47	AA	F	30	0	0		0	2	1991	0	0	0	2	66	1.7	65	82
34	unemployed	GB1022	54	AA	F	34	0	0	weed	0	3	2003	1	10	MTx	1.85	75	1.42	73	77
35	student	JB9786	18	AA	M	23	0	0		0	2	2008	1	0	0	3.8	93	3.13	83	82
36	student	JB3763	29	AA	F	40.3	0	0		0	1	2002	0	0	0	3.24	82	2.8	89	86
37	truck driver	KB6965	32	AA	M	30	1	0		10	2	2007	1	60	0					
38	unemployed	KB8507	57	AA	F	28	0	0		0		2005	1			1.6	56	1.5	72	96
39	clerk	KFB4122	58	?	F	25	0	1		15	1	2000	0	0	0					
40	State employee	LB3994	57	AA	F	32	0	1		5	2	2001	0	0	0	2.29	79	1.82	84	79
41	SSI	MB1293?	34	AA	F	40.2	1	0		50	0					2.9	80	2.32	82	80
42	cashier	MB9785	24	AA	F	19.3	0	0		0	4	1994	1	40 QD	MTX 10	0.9	24	0.8	25	88
43	retired	RB4709	62	AA	M	30	0	0		0	1	2005	0	0	0	2.05	45	1.34	42	66
44	unemployed	RB7477	43	AA	M	22	1	0	cocaine	20	2	1995	0	0	0	4.3	83	3.29	85	77

etc.:

. . . and much more . . . .

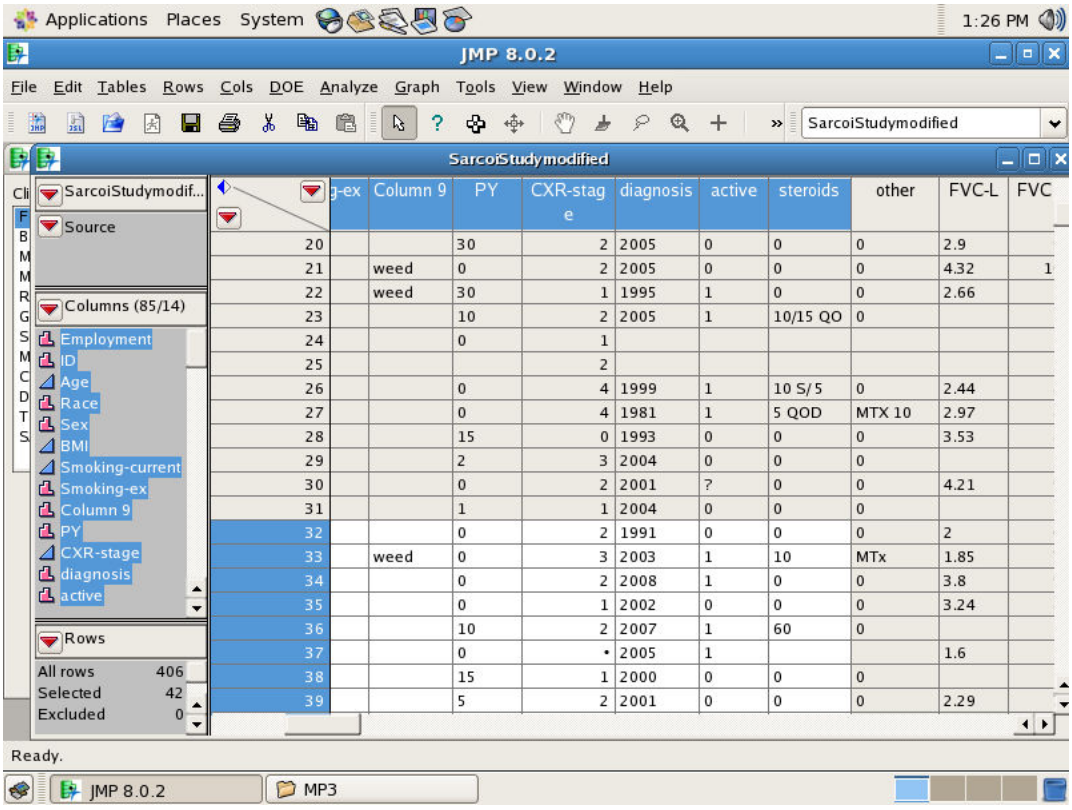
CLEANSING THE DATA:

The JMP method for correcting data types is simple point-and-click, but the Base SAS method is convoluted by assignment statements, drop statements, put/input statements, etc. In EXCEL of course you can also point-and-click most of the necessary cleansing, but you have to know what you want because EXCEL won't tell you where the data might be wrong for your analysis.

The JMP solution for parsing columns ("substring extraction"), as will be needed in the "steroids" column below (where steroid quantities and the steroid descriptions are mixed) is to send the data to SAS for subsetting, because JMP does not have a parser. SAS automatically returns the results to JMP for JMP's analysis after the parsing is done by SAS, but if you don't have SAS to help you on your JMP platform you'll have to pre-process your columns with the EXCEL "mid" or "text-to-column" functions before it goes to JMP—or you can do it all manually of course.

In BASE SAS you use PROC IMPORT to bring the Excel data into SAS, then you use the DATA Step with various SAS statement lines to assign new numeric variables to existing character variables (or vice versa), then you drop your old values. Perhaps you add a SAS statement or two to organize the presentation order of the fields, or to document your logic, and then you run the SAS job and you inspect the SAS LOG for any oddities. SAS/Stat Studio makes the process much easier with its JMP-like point-and-click view to the data, and SAS Enterprise Guide makes the process easier by generating much of the DATA Step for you, but those SAS products are extra items beyond the central product, whereas with JMP we've already got the functionality available to us.

Here is the JMP reading of the EXCEL data:

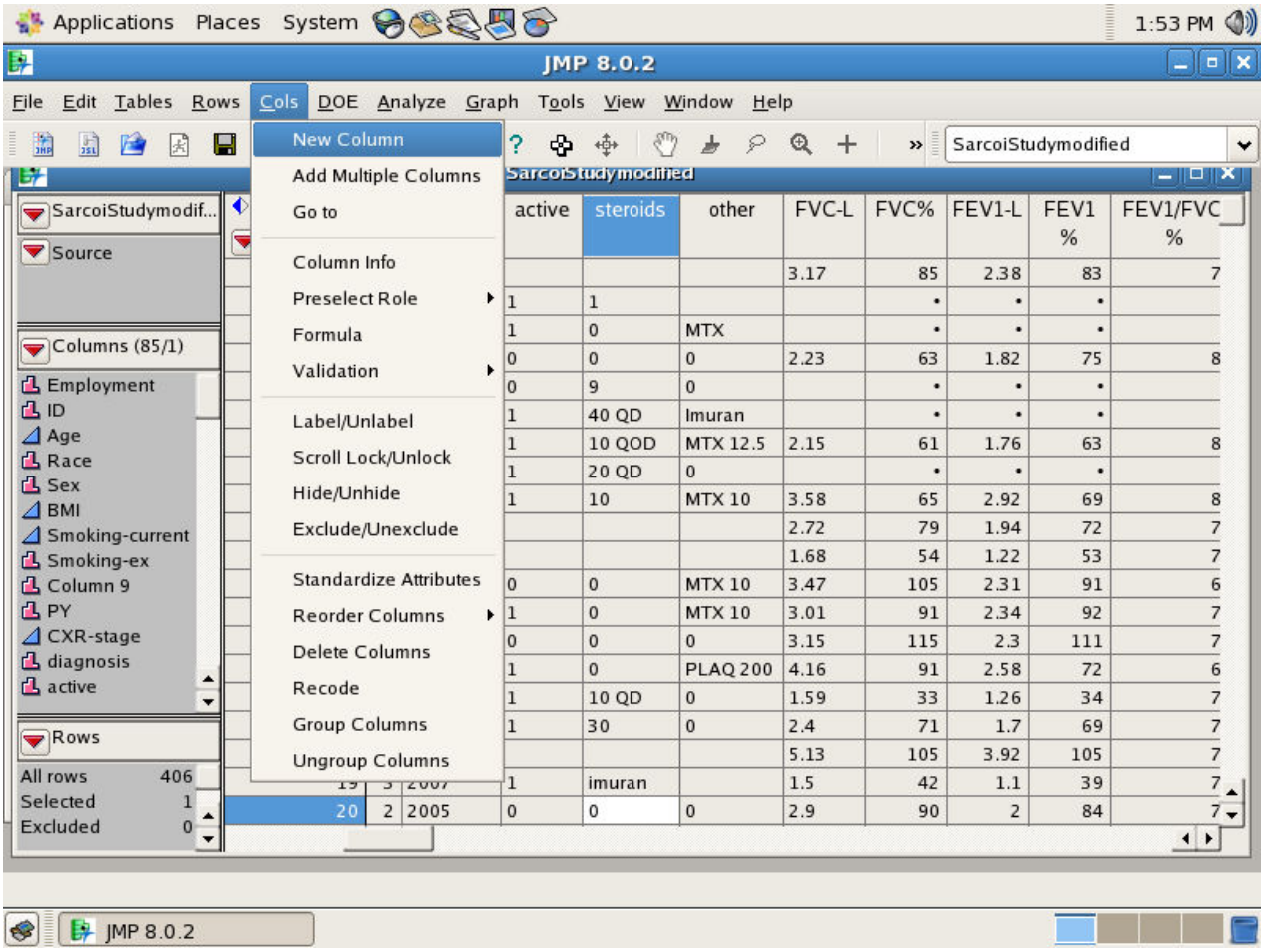


	g-ex	Column 9	PY	CXR-stage	diagnosis	active	steroids	other	FVC-L	FVC
20			30	2	2005	0	0	0	2.9	
21		weed	0	2	2005	0	0	0	4.32	1
22		weed	30	1	1995	1	0	0	2.66	
23			10	2	2005	1	10/15 QO	0		
24			0	1						
25				2						
26			0	4	1999	1	10 S/ 5	0	2.44	
27			0	4	1981	1	5 QOD	MTX 10	2.97	
28			15	0	1993	0	0	0	3.53	
29			2	3	2004	0	0	0		
30			0	2	2001	?	0	0	4.21	
31			1	1	2004	0	0	0		
32			0	2	1991	0	0	0	2	
33		weed	0	3	2003	1	10	MTx	1.85	
34			0	2	2008	1	0	0	3.8	
35			0	1	2002	0	0	0	3.24	
36			10	2	2007	1	60	0		
37			0	•	2005	1			1.6	
38			15	1	2000	0	0	0		
39			5	2	2001	0	0	0	2.29	

JMP reading of the EXCEL data (cont.):

As we see, some variables, like PY ("Pack-Years") are read-in as character by default, but we want this to be continuous so that numeric analysis is possible with this field.

Other variables, like "Column 9" are vague. Still other fields, like "Steroids" (for "Steroid Usage") and "Other" seem to be a mix of both nominal and ordinal meanings, so we should get some clarification from the doctor before we use these fields in a model. Some cases contain lots of missing values, so our doctor should comment about them to us before we complete the analysis. Here is an example of using JMP to separate the Steroids column into two columns (one for the metering, and another for the labeling):



JMP lets us by point-and-click to add new columns and re-order them for desired presentation in the data. We can only generate *p-value*s on numeric columns, so it is important that all candidates for such analysis be properly classified as continuous. In simple situations, like with "FVC-L" in the above table, we need only right-click on the column heading and specify that we want numeric orientation (any spurious characters will cause the cell to convert to "missing" [“.”]) of the data instead of character orientation.

UNDERSTANDING THE STATISTICS:

The doctor's initial request was that a table of simple statistics and *p-value*s be produced to describe the data under analysis when SMOKERS are correlated to NON-SMOKERS. We would like to model the treatment success factors to make breathing easier for sarcoidosis patients, perhaps incorporating nuances like "substance abuse" or "race" or "quit smoking" to discover variations in success along the way. We would then proceed to understand success factors in the treatment of sarcoidosis patients, many of whom are smokers. The analyst recalled that the *p-value* is a measure to establish the likelihood of a weak statistical correlation. Formally "the *p-value* of the correlation *r* is the probability of obtaining a Student's *t* statistic greater in absolute value than the absolute value of the observed statistic *t*".[1] In other words, the *p-value* is "the cutoff probability for declaring an insignificant difference."[2] In the JMP Manual itself[3], "*p* [is] the attributed probability." Sometimes the *p-values*, or simply "*p*," are called "Pearson correlation statistics," named after the English statistician Carl Pearson (who got nicknamed "Karl" when he went to school in Heidelberg in the 1880s).

The *p-value* arises out of the univariate t-test, and SAS Software® offers PROC TTEST and PROC MULTTEST to generate *p-value*s by default. JMP Software® generates *p-value*s not by default but as an option. PROC TTEST only lets you analyze one variable at a time to generate the *p-value*, but PROC MULTTEST allows any number of variables to be considered in a single pass of the data. JMP Software® generates default statistics for multiple variables in a single step, but you have to individually derive the *p-value* as an option for each variable after the initial JMP display is done.

The JMP manual is forthright in commenting that JMP is superb at deriving statistics, but JMP does not purport to be a reporting tool for those statistics. That is a reason why JMP allows us to dip into SAS for help in processing our data, and a quick-and-clean solution to the problem presenting a custom report of *p-value*s and other statistics for an unknown number of variables is to call SAS from JMP and execute a PROC MULTTEST program[4] to generate the *p-value*s for all of our numeric variables. Here is an example of the desired display format without regard to the qualifying *p-value*s (see next page):

[1] From Help Screen in SAS Software
[2] From Help Screen is SAS Software
[3] From "JMP, Statistics and Graphics Guide Volume 1, JMP8," p 679"
[4] See Appendix A

Display format of numeric variables (cont.)

Display of Variables in Sarcoidosis Study

Obs	_NAME_	_NMISS_	MIN	MAX	MEAN	STD	_NNONMISS_	MIN	MAX	MEAN	STD	PVALUE
1	ACE_Level	55	0.00	195.00	63.036	46.36	162	0.00	540.00	65.741	62.02	0.76727
2	Age	100	22.00	82.00	44.430	9.72	273	18.00	85.00	48.403	12.22	0.00362
3	Alk_phos	14	63.00	1210.00	185.643	295.98	42	49.00	576.00	146.024	122.67	0.47955
4	BMI	93	17.80	48.20	29.915	7.22	274	19.20	70.00	33.444	8.88	0.00059
5	Borg_1min	45	0.00	10.00	2.444	2.18	178	0.00	7.00	1.498	1.62	0.00133
6	Borg_6min	45	0.00	10.00	4.800	2.31	178	0.00	10.00	4.293	2.46	0.21239
7	CXR_stage	74	0.00	4.00	1.676	1.16	218	0.00	4.00	1.899	1.33	0.19969
8	Calcium	75	0.00	104.00	9.475	11.39	206	0.00	12.80	8.493	2.77	0.25125
9	CathPHTN	0	.	.	.	.	2	1.00	1.00	1.000	0.00	1.00000
10	Change_in_sat	45	-2.00	9.00	2.444	1.97	179	-1.00	22.00	3.659	4.05	0.05195
11	DLCO_L_mmHg_min	70	2.21	38.10	15.919	7.61	230	0.77	34.30	16.271	6.08	0.69023
12	DLCO__	70	21.00	124.00	61.857	22.99	231	2.30	122.00	64.920	22.18	0.31649
13	DSP	33	102.90	601.60	417.983	117.47	148	102.30	662.40	376.120	108.69	0.05023
14	Echo_EF	35	20.00	80.00	57.429	8.69	134	15.00	80.00	56.396	9.70	0.56758
15	Echo_PHTN	30	0.00	2.00	0.800	0.92	117	0.00	75.00	3.239	9.91	0.18126
16	Eye	4	1.00	1.00	1.000	0.00	6	0.00	1.00	0.500	0.55	0.11143
17	FEV1_	71	28.00	127.00	76.479	23.45	237	21.00	127.00	74.654	21.50	0.53949
18	FEV1_FVC__	71	41.00	87.00	74.859	9.05	238	3.45	99.00	76.650	11.10	0.21533
19	FEV1_L	71	0.74	5.30	2.370	1.03	236	0.66	4.07	2.074	0.71	0.00648
20	FM_Pain	1	1.00	1.00	1.000	.	5	0.00	1.00	0.600	0.55	0.54147
21	FVC_	71	40.00	127.00	77.521	20.06	238	24.00	122.00	74.349	19.34	0.23011
22	FVC_L	71	1.09	6.76	3.072	1.22	235	0.90	5.55	2.688	0.86	0.00336
23	Hep_C	14	0.00	1.00	0.214	0.43	48	0.00	1.00	0.063	0.24	0.09385
24	Lung	10	1.00	1.00	1.000	0.00	22	0.00	1.00	0.909	0.29	0.34066
25	Lung_Bx	11	1.00	6.00	2.273	2.05	27	0.00	4.00	0.852	0.91	0.00502
26	Lymph_Node_Bx	5	0.00	1.00	0.400	0.55	8	0.00	1.00	0.500	0.53	0.75112
27	Nodes	8	0.00	1.00	0.875	0.35	19	0.00	1.00	0.789	0.42	0.61781
28	Ox_sat	63	72.00	100.00	97.365	3.63	220	73.00	101.00	96.977	3.58	0.45029
29	PY	90	0.00	60.00	12.806	13.69	244	0.00	80.00	6.697	11.08	0.00004
30	Skin_Bx	4	0.00	1.00	0.750	0.50	17	0.00	1.00	0.882	0.33	0.52054
31	Smoking_ex	94	0.00	12006.00	148.862	1252.67	213	0.00	20008.00	206.516	1439.09	0.73695
32	Spleen	2	1.00	1.00	1.000	0.00	4	0.00	1.00	0.750	0.50	0.54147
33	TLC	70	2.07	10.17	4.814	1.55	225	1.43	68.00	4.817	5.27	0.99638
34	TLC_	70	54.00	140.00	86.471	17.32	226	3.80	137.00	80.123	19.22	0.01408
35	_FREQ_	100	100.00	100.00	100.000	100.00	277	277.00	277.00	277.000	277.00	.
36	_MW_dist	45	105.00	720.00	436.111	111.75	180	115.00	690.00	402.056	118.13	0.08186
37	_MW_dist_	44	22.80	104.00	75.098	16.04	176	21.00	153.00	76.063	20.90	0.77528
38	_MW_ox_at_1	45	94.00	100.00	97.444	1.63	181	90.00	100.00	97.569	1.95	0.69249
39	_MW_ox_at_6	45	88.00	100.00	95.200	2.75	181	76.00	100.00	94.072	4.78	0.12983

When we qualify the variables by their *p-value*s, the table of eligible variables is much smaller (see next page):

When we qualify the variables by their *p-value*s . . . (from previous page):

Display of Qualifying modeling Variables in Sarcoidosis Study

Obs	_NAME_	_N_OBS_	_MIN_	_MAX_	_MEAN_	_STD_	_NMISS_	MIN_	MAX_	MEAN_	STD_	P_VALUE
1	Age	100	22.00	82.00	44.430	9.724	273	18.00	85.00	48.403	12.222	0.003622
2	BMI	93	17.80	48.20	29.915	7.222	274	19.20	70.00	33.444	8.875	0.000594
3	Borg_1min	45	0.00	10.00	2.444	2.177	178	0.00	7.00	1.498	1.619	0.001329
4	FEV1_L	71	0.74	5.30	2.370	1.033	236	0.66	4.07	2.074	0.714	0.006477
5	FVC_L	71	1.09	6.76	3.072	1.224	235	0.90	5.55	2.688	0.864	0.003357
6	Lung_Bx	11	1.00	6.00	2.273	2.054	27	0.00	4.00	0.852	0.907	0.005022
7	PY	90	0.00	60.00	12.806	13.691	244	0.00	80.00	6.697	11.079	0.000037
8	TLC_	70	54.00	140.00	86.471	17.321	226	3.80	137.00	80.123	19.219	0.014084
9	_FREQ_	100	100.00	100.00	100.000	100.000	277	277.00	277.00	277.000	277.000	.

MODELING THE INFORMATION:

Sarcoidosis is a “dreaded-disease,” wherein open sores form internally on body organs. How about a messy gash on the pancreas? Maybe a puss-filled boil on the kidney? Maybe an infected rupture in the salivary glands? It’s like that, except that it’s not from an injury and it can’t be healed. It can only be treated: Sarcoidosis can be managed by medication and surgery, and even such treatments can be themselves experimental and dangerous. Hence research studies like this very paper arise to quantify and qualify the potential for improving the affected lifestyles.

Your comments and questions are valued and encouraged. Contact the author at:

Name	Jim Woods
Enterprise	Wayne State University
Address	1221 Hampton Road
City, State ZIP	Detroit, MI 48236-1362
Phone:	313-884-7354, 313-320-7983
Fax:	877-350-3612
E-mail:	jim.woods@gljug.com
Web:	wkrell@med.wayne.edu

APPENDIX A: SAS Program to Generate *P-value*s and other statistics and report the results from Excel source data

```
PROC IMPORT
   OUT=SASUSER.SarcoidosisData
   DATAFILE= "C:\ . . .\SarcoiStudymodified.xls"
   DBMS=EXCEL REPLACE;
   RANGE="SarcoidosisData";
   GETNAMES=YES;
   MIXED=NO;
   SCANTEXT=YES;
   USEDATE=YES;
   SCANTIME=YES;
RUN;
proc multtest
   noprint
   out=work.sarcoidosispvalues(keep=_var_ raw_p rename=(raw_p=P_VALUE))
   data=sasuser.sarcoidosisdata(drop=f4: f5:);
   class smoking_current smoking_ex;
   test mean(_numeric_);
run;
proc summary
   missing
   data=sasuser.sarcoidosisdata;
   class smoking_current smoking_ex;
   var _numeric_;
   output out=work.sarcoidosistests;
run;
data sasuser.sarcoidosistestdata(drop=_type_);
   set work.sarcoidosistests;
   where smoking_current ne .;
run;
proc transpose data=sasuser.sarcoidosistestdata out=tests;
   by smoking_current;
   id _stat_;
run;
data
   smokers
      (rename=
      (N=N_SMOKERS
       MIN=MIN_SMOKERS
       MAX=MAX_SMOKERS
       STD=STD_SMOKERS
       MEAN=MEAN_SMOKERS
      ))
   nonsmokers
      (rename=
      (N=N_NONSMOKERS
       MIN=MIN_NONSMOKERS
       MAX=MAX_NONSMOKERS
       STD=STD_NONSMOKERS
       MEAN=MEAN_NONSMOKERS
      ))
   ;
   set tests;
   drop _label_ smoking_current;
   if smoking_current=0 then output nonsmokers;
   else output smokers;
run;
proc sql;
   create table sasuser.sarcoidosisresults(drop=_var_)
   as select *
   from work.smokers a
   left join work.nonsmokers b
   on a._name_=b._name_
   left join work.sarcoidosispvalues c
   on a._name_=c._var_
   ;
quit;
```