Paper 228-2008

Horizon Estimation Using SAS® Software

Gary A. Foster, St. Joseph's Healthcare, Hamilton ON, Canada
Charles H. Goldsmith, St. Joseph's Healthcare & McMaster University, Hamilton ON, Canada

ABSTRACT

When planning a clinical trial, or when collecting information to include in a meta-analysis, an investigator typically searches the available literature to get the most up-to-date bibliographic information available. Such a literature search, or systematic review, is usually performed by searching databases, but it may also involve a search of books, journals, and websites. Common database searches for medical researchers are the Cochrane Library, Medline, EMBASE, PsycINFO, and CINAHL, to name a few.

Typically the goal of the review is to find as many relevant articles as possible, but such a review need not be limited to articles. Here we present an example where the search is for relevant journals rather than articles. It is quite likely that a review will miss some, if not many, articles or journals. Some journals could be missed because the review was limited; in its key words, to one or a small number of languages, to a particular time frame, or to a small number of databases. Knowing that it is unlikely that a review will find all relevant journals, it is important for a researcher to estimate the number of journals that may have been missed. Adding this estimate to the total number of journals identified by the review defines the Horizon. If a review identifies a small proportion of journals relative to the Horizon, the research project being planned may be based on inadequate information.

This paper presents a method based on Poisson regression modeling to estimate the number of journals that have been missed by a review. A SAS solution to this problem provides the results of a worked example with 4 databases.

INTRODUCTION

Suppose a researcher is planning a clinical trial to test a new drug for the treatment of a kidney condition. As part of her plan she would like to identify all journals in the nephrology literature that address treatment issues for that condition. She begins by searching the Cochrane database and finds 102 relevant journals. Next she searches the Plus database and comes up with 241 journals, and finds 444 journals in Medline. Finally she searches the Renal database and finds 275 journals. The data from these searches are then organized into the NEPH data set, shown as a listing in Table 1.

TABLE 1: Listing of the NEPH data set containing the nephrology systematic review results

obs	A	B	C	D	count
1	1	1	1	1	77
2	1	1	1	0	11
3	1	1	0	1	1
4	1	1	0	0	3
5	1	0	1	1	3
6	1	0	1	0	7
7	1	0	0	1	0
8	1	0	0	0	0
9	0	1	1	1	88
10	0	1	1	0	61
11	0	1	0	1	0
12	0	1	0	0	0
13	0	0	1	1	88
14	0	0	1	0	109
15	0	0	0	1	18
16	0	0	0	0	.

The NEPH data set has been organized into five columns, where the first four (excluding obs) represent the four databases (A = Cochrane, B = Plus, C = Medline, D = Renal) and the rightmost column the number of journals found. Each of the first four columns contains either a 1 or a 0 in every row. In total there will be 2^k rows, one row for each possible combination of k databases being searched. If n1 articles were common to all four databases then there would be a 1 under each database name and n1 under COUNT (see row 1 for an example when n1 equals 77) in that row. Similarly, if n10 journals were common to the Plus and Medline searches but did not turn up in the Cochrane or Renal database searches then n10 would be in the COUNT column on the row where there is a 0 under Cochrane and Renal and a 1 under PLUS and Medline (see row 10 where n10 equals 61).

The number of journals found in each search can be determined from this table by summing the counts from the relevant rows. For example, the total number of journals found by the search of the Cochrane database is determined by summing the numbers in the count column across the first 8 rows (all the rows with a 1 under Cochrane). Similarly the number of journals found by the PLUS search would equal the sum of the numbers in the count column across the first four rows and rows 9 through 12, wherever there is a 1 in the Plus column. Table 2 summarizes the number of journals found by each database search by itself.

TABLE 2: Number of journals identified in each database separately

Cochrane	(A)	102
Plus	(B)	241
Medline	(C)	444
Renal	(D)	275

In total the review found 466 different journals, 77 that were identified in each of the four databases. No journals were found exclusively by the Cochrane or Plus databases. In row 15 we see that 18 journals were found in Renal that were not found in any of the others. The last row of Table 1 reminds us that there is an unknown number of journals that were not identified by any of the four searches. If we were able to estimate this unknown number it would allow us to estimate the total number of journals published on the topic, the Horizon. Once we have an estimate of the Horizon we could compute a measure of how well the review performed.

Since the task is to estimate the number of missed journals we will use Poisson regression as it is often used to model frequency or count data.

The Horizon macro will provide an estimate of the Horizon for this nephrology example. Before we pass the data to the macro the data set needs to include all possible interaction terms for potential use in the models. To create each interaction term, simply set a new variable to the product of the appropriate main effect terms as shown in the data step example below. Notice that the four way interaction term is not required as that term would over saturate the model.

```
data neph;
   set neph;
   ab  = a*b;
   ac  = a*c;
   ad  = a*d;
   bc  = b*c;
   bd  = b*d;
   cd  = c*d;
   abc = a*b*c;
   abd = a*b*d;
   acd = a*c*d;
   bcd = b*c*d;
run;
```

Missing journals and Horizon estimate from the first 2 databases searched

Suppose the researcher had stopped the review after completing searches of the Cochrane (A) and Plus (B) databases. By collapsing the data in Table 1 for these two databases, a summary of the journals found in this review is provided in Table 3. If we sum the numbers in the count column we see that a total of 251 journals were found in these two searches.

TABLE 3: Nephrology journals found in Cochrane (A) and Plus (B) database searches

Obs	A	B	count
1	1	1	92
2	1	0	10
3	0	1	149
4	0	0	Unknown

We are now ready to run the Poisson regression analysis with COUNT as the dependent variable. For this table there is one model, the saturated model that has 3 terms: the grand mean, the effect of A (Cochrane) and the effect of B (Plus). Since there are 3 parameters to estimate, and there are exactly 3 rows of data with count information, the model fit is exact, and there are no degrees of freedom for estimating error. The model saturates the rows, so the fits will be the actual data and the residuals will all be zeroes. An estimate of the intercept (coefficient estimate), along with other details of this model, are shown in Table 6 in the row for Model #1.

When terms in the model are set to zero (A = 0 and B = 0) the best estimate for the number of missed journals is the intercept. This model fits the data perfectly and the missing cell estimate is 17, and its 95% CI is [8,32]. By adding the number of known articles (i.e., 251) to these numbers we obtain the Horizon estimate and its 95% CI: 268 and [259, 283] respectively. Our review found 251 out of a Horizon of 268 journals, a relatively good result with an estimated 94% hit rate.

Since the additive, or saturated model in this case is the only model we could fit to the data, that takes into account both databases and the grand mean, we may be mislead into believing that this model is a correct characterization of the Cochrane and Plus databases together. However, since the missing cell estimate and the Horizon estimate depend on the model fitted, we are not in a position to evaluate whether that is the case or not. We can however find out whether the additive assumption is realistic by including the results of one or more additional database searches to the model.

Missing journals and Horizon estimate from the first 3 databases searched

If the results of the third database are included with the first two, the findings can be summarized as shown in Table 4. For these three searches a total of 448 journals were identified.

TABLE 4: Nephrology journals found in Cochrane, Plus and Medline searches

Obs	A	B	C	count
1	1	1	1	88
2	1	1	0	4
3	1	0	1	10
4	1	0	0	0
5	0	1	1	149
6	0	1	0	0
7	0	0	1	197
8	0	0	0	Unknown

We now use these data, along with all 2-factor interaction terms to find the best model from all the various models available to be fitted. We first consider the model what contains the 3 main effects and later consider other models that contain some or all of the 2-factor interactions. The results of these models are shown in Models# 2 through 9 in Table 6.

The best model, when interaction terms are entered singly into the model, is Model # 3 because it has the lowest deviance at 8.14. This model has 3 main effects plus the AB (Cochrane*Plus) interaction term which indicates that these two databases interact. This fact leads us to reject the simple two database model specified earlier. If we next inspect the 1 df models that include the AB interaction term, we find that the model with the AC (Cochrane*Medline) interaction term has the smallest deviance equal to 0.84. The difference between the deviances of these two models equals 7.30 which, when judged by a 1 df Chi-square, is a large and statistically significant difference at the 5% level. This suggests that the AC interaction term is also important.

If we finally consider the saturated model, we see that the model fits the data exactly as the deviance equals 0 and there are 0 df. For an improvement of 0.84 in the deviance at the cost of using 1 df this does not appear to be a better model than the best 1 df model identified above. For this reason the authors would select Model # 6 to estimate the Horizon. That model provides a missing cell estimate of 1 with a 95% CI of [0, 1] so the Horizon

estimate is 449 and its 95% CI is [448, 449]. The closeness to this Horizon estimate is 99.8% so the model looks very good. However, this 3 database review identified 448 known journals. If the results of the renal database search are included in the review, the number of known journals increases to 466. For this reason the four database review will be evaluated next.

Missing journals and Horizon estimate from all 4 database searches

The organization of the data set for this analysis is shown in Table 1 and the results of all the models are shown in Table 6 starting at Model # 10. Model #10 shows the coefficient estimate and its standard error equal 1.4209 and 0.2322 respectively for the main effects model (no interaction terms). To estimate the number of missed journals, we raise the mathematical constant e to the power of the coefficient estimate to obtain 4.14 which, when rounded up to the next integer, yields an estimate of 5. Similarly, the lower bound of the 95% confidence interval (CI) equals exp(1.4209-1.96*(0.2322)) and the upper 95% CI bound equals exp(1.4209 +1.96*(0.2322)). These computations provide estimates of the missing number equal to 5 with 95% CI equal to [2,7]. To obtain the Horizon estimate along with its 95% CI add the total number of known articles (466) to these numbers to arrive at 471 and [468, 473]. Finally, by dividing the total number of known articles by the Horizon estimate (466/471) we find that the researcher has identified about 99% of the estimated number of Horizon journals. This appears to be an impressively high percentage, but the main effects model is quite poor as the deviance estimate equals 175.87 with 10 df.

As can be seen in Table 6, many four database models were considered with some being better than others. To determine the best model we add interaction terms to the main effects model and compare the deviance estimates and degrees of freedom amongst these models. The best model with 9 df is model #11 which introduces the AB interaction term. It has a deviance equal to 86.52 which is significantly better than the main effects model with 10 degrees of freedom. (Recall that this was also an important interaction term when only three searches were evaluated.) Next we consider models with 8 df that include the AB term (Model #s 17 – 21) . Of these models Model #18 has the smallest deviance at 60.98, with the AD interaction term being added to the model. Of all models with 7 df that include the AB and AD interaction terms, Model #36 has the smallest deviance and this deviance is significantly better than the best model with 8 df. There are three models with 6 df that include the AB, AD, and BC interaction terms (Model #s 52, 58, 59). The best amongst these is Model # 58 which has a deviance equal to 26.29. Again, this model is significantly better than the best model with 7 df. As we attempt to progress further using this approach it becomes evident that the models do not improve significantly and in some cases they become unstable. Now we must consider additional models building on Model # 58 but with the appropriate three-way interaction terms. These models are shown as Models # 89 through 95. Upon considering these models, it becomes apparent that all three way interaction terms add significantly to the prediction of the missing cell estimate.

Of the many models considered, we feel the best is Model # 95. The estimate of the unknown number of journals is 21 with a 95% CI [12, 36]. Since there are 466 known journals, these numbers allow us to compute the Horizon estimate to be 487 with 95% CI [478, 502]. Thus, the investigator has identified about 96% (466/487) of the Horizon journals. This model is superior to all other models considered as its deviance equals 2.57 with 3 df. Table 5 provides a comparison of the best model from reviews including 2, 3, and 4 databases.

TABLE 5: Summary of best model by number of databases in review

Number of databases searched	'Best' model from those considered*	Number of known journals	Horizon estimate with 95% CI	Journals found as % of Horizon	Number of additional journals found
1	A	102			
2	A B	251	268 [259, 283]	94%	149
3	A B C AB AC	448	449 [448, 449]	100%	197
4	A B C D AB AD BC BD ABC ABD ACD	466	487 [478, 502]	96%	18

* A=Cochrane, B=Plus, C=Medline, D=Renal

The macro computes estimates and 95% CIs of the unknown number of articles and the Horizon and lists them in the output along with deviance and other information about each model evaluated. The modeling section of the output is shown in Table 6 below.

TABLE 6: Estimates of unknown number and Horizon for models fitted to 2, 3, and 4 database searches

Nephrology Study: Estimating the Horizon
Comparison of Various Models

Model#	Model evaluated	Missing cell Estimate	Missing cell Lower 95% CI	Missing cell Upper 95% CI	# of known journals	Coefficient estimate	Coefficient Lower 95% CI	Coefficient Upper 95% CI	Horizon estimate	Horizon Lower 95% CI	Horizon Upper 95% CI	Known journals as % of Horizon	Deviance	DF
1	a b	17	8	32	251	2.7847	2.1127	3.4568	268	259	283	94%	0.00	0
2	a b c	3	0	7	448	0.8086	-0.1868	1.8039	451	448	455	99%	89.32	3
3	a b c ab	4	1	9	448	1.1601	0.1624	2.1578	452	449	457	99%	8.14	2
4	a b c ac	1	0	1	448	-19.8452	-20.8555	-18.8349	449	448	449	100%	80.91	2
5	a b c bc	1	0	1	448	-18.4717	-19.4705	-17.4729	449	448	449	100%	87.88	2
6	a b c ab ac	1	0	1	448	-19.4139	-20.4234	-18.4044	449	448	449	100%	0.84	1
7	a b c ab bc	1	0	1	448	-16.7125	-17.7106	-15.7145	449	448	449	100%	7.81	1
8	a b c ac bc	0	0	0	448	-42.7726	-43.7951	-41.7501	448	448	448	100%	75.92	1
9	a b c ab ac bc	0	0	0	448	-40.3186	-41.3303	-39.3069	448	448	448	100%	0.00	0
10	a b c d	5	2	7	466	1.4209	0.9658	1.8760	471	468	473	99%	175.87	10
11	a b c d ab	6	3	9	466	1.6713	1.2156	2.1271	472	469	475	99%	86.53	9
12	a b c d ac	5	2	8	466	1.5213	1.0253	2.0172	471	468	474	99%	175.11	9
13	a b c d ad	5	3	8	466	1.5835	1.1261	2.0408	471	469	474	99%	150.66	9
14	a b c d bc	9	5	15	466	2.1845	1.6899	2.6791	475	471	481	98%	158.07	9
15	a b c d bd	6	3	9	466	1.7174	1.2517	2.1832	472	469	475	99%	153.41	9
16	a b c d cd	2	0	6	466	0.6054	-0.5388	1.7497	468	466	472	100%	172.30	9
17	a b c d ab ac	6	3	10	466	1.7879	1.2906	2.2851	472	469	476	99%	85.56	8
18	a b c d ab ad	7	3	10	466	1.8434	1.3847	2.3020	473	469	476	99%	60.98	8
19	a b c d ab bc	13	7	21	466	2.5149	2.0163	3.0134	479	473	487	97%	66.44	8
20	a b c d ab bd	8	4	12	466	1.9887	1.5202	2.4573	474	470	478	98%	63.40	8
21	a b c d ab cd	3	0	9	466	0.9558	-0.1904	2.1021	469	466	475	99%	83.97	8
22	a b c d ac ad	6	3	9	466	1.6940	1.1953	2.1927	472	469	475	99%	149.77	8
23	a b c d ac bc	11	6	19	466	2.3631	1.8224	2.9038	477	472	485	98%	156.31	8
24	a b c d ac bd	7	3	11	466	1.8388	1.3308	2.3469	473	469	477	99%	152.40	8
25	a b c d ac cd	3	0	7	466	0.7113	-0.4716	1.8942	469	466	473	99%	171.90	8
26	a b c d ad bc	12	6	19	466	2.3985	1.8987	2.8984	478	472	485	97%	131.45	8
27	a b c d ad bd	7	4	11	466	1.8935	1.4239	2.3631	473	470	477	99%	127.79	8
28	a b c d ad cd	3	0	8	466	0.8275	-0.3195	1.9744	469	466	474	99%	147.73	8
29	a b c d bc bd	14	8	23	466	2.5977	2.0815	3.1140	480	474	489	97%	132.73	8
30	a b c d bc cd	7	1	22	466	1.8187	0.5761	3.0614	473	467	488	99%	157.62	8
31	a b c d bd cd	3	0	9	466	1.0168	-0.1392	2.1728	469	466	475	99%	151.04	8
32	a b c d ab ac ad	8	4	12	466	1.9743	1.4733	2.4753	474	470	478	98%	59.81	7
33	a b c d ab ac bc	12	7	20	466	2.4716	1.9573	2.9859	478	473	486	97%	65.76	7
34	a b c d ab ac bd	9	5	15	466	2.1359	1.6232	2.6486	475	471	481	98%	62.02	7
35	a b c d ab ac cd	3	0	10	466	1.0942	-0.1034	2.2919	469	466	476	99%	83.45	7
36	a b c d ab ad bc	16	9	27	466	2.7663	2.2592	3.2734	482	475	493	97%	38.73	7
37	a b c d ab ad bd	8	4	12	466	2.0147	1.5464	2.4830	474	470	478	98%	51.48	7
38	a b c d ab ad cd	4	1	11	466	1.2196	0.0688	2.3705	470	467	477	99%	59.16	7
39	a b c d ab bc bd	21	12	36	466	3.0246	2.4912	3.5581	487	478	502	96%	38.69	7
40	a b c d ab bc cd	20	4	88	466	2.9703	1.4633	4.4773	486	470	554	96%	66.07	7
41	a b c d ab bd cd	5	1	14	466	1.4589	0.2925	2.6253	471	467	480	99%	62.20	7
42	a b c d ac ad bc	14	7	24	466	2.6159	2.0660	3.1658	480	473	490	97%	129.13	7
43	a b c d ac ad bd	8	4	13	466	2.0311	1.5179	2.5442	474	470	479	98%	126.55	7
44	a b c d ac ad cd	3	0	8	466	0.9172	-0.2358	2.0702	469	466	474	99%	146.49	7
45	a b c d ac bc bd	18	9	32	466	2.8714	2.2948	3.4481	484	475	498	96%	129.65	7
46	a b c d ac bc cd	10	2	44	466	2.2702	0.7603	3.7800	476	468	510	98%	156.30	7

a = Cochrane, b = PLUS, c = Medline, d = Renal

Nephrology Study: Estimating the Horizon
Comparison of Various Models

Model#	Model evaluated	Missing cell Estimate	Missing cell Lower 95% CI	Missing cell Upper 95% CI	# of known journals	Coefficient estimate	Coefficient Lower 95% CI	Coefficient Upper 95% CI	Horizon estimate	Horizon Lower 95% CI	Horizon Upper 95% CI	Known journals as % of Horizon	Deviance	DF
47	a b c d ac bd cd	4	0	11	466	1.1642	-0.0483	2.3767	470	466	477	99%	150.50	7
48	a b c d ad bc bd	18	10	31	466	2.8720	2.3414	3.4025	484	476	497	96%	104.66	7
49	a b c d ad bc cd	12	2	43	466	2.4115	1.0639	3.7590	478	468	509	97%	131.45	7
50	a b c d ad bd cd	4	1	12	466	1.2927	0.1285	2.4568	470	467	478	99%	126.18	7
51	a b c d bc bd cd	9	2	27	466	2.0899	0.9058	3.2740	475	468	493	98%	131.62	7
52	a b c d ab ac ad bc	16	9	26	466	2.7255	2.2020	3.2490	482	475	492	97%	38.21	6
53	a b c d ab ac ad bd	9	5	15	466	2.1647	1.6521	2.6773	475	471	481	98%	50.06	6
54	a b c d ab ac ad cd	4	1	13	466	1.3322	0.1722	2.4922	470	467	479	99%	57.80	6
55	a b c d ab ac bc bd	20	11	35	466	2.9864	2.4354	3.5375	486	477	501	96%	38.33	6
56	a b c d ab ac bc cd	16	3	77	466	2.7622	1.1897	4.3347	482	469	543	97%	65.62	6
57	a b c d ab ac bd cd	6	1	18	466	1.6598	0.4433	2.8762	472	467	484	99%	61.16	6
58	a b c d ab ad bc bd	22	12	37	466	3.0681	2.5336	3.6026	488	478	503	95%	26.29	6
59	a b c d ab ad bc cd	37	7	187	466	3.6086	1.9897	5.2275	503	473	653	93%	37.67	6
60	a b c d ab ad bd cd	5	1	15	466	1.5065	0.3393	2.6737	471	467	481	99%	50.39	6
61	a b c d ab bc bd cd	33	8	130	466	3.4805	2.0994	4.8615	499	474	596	93%	38.24	6
62	a b c d ac ad bc bd	24	12	42	466	3.1460	2.5609	3.7311	490	478	508	95%	101.46	6
63	a b c d ac ad bc cd	18	3	86	466	2.8827	1.3132	4.4521	484	469	552	96%	129.00	6
64	a b c d ac ad bd cd	5	1	14	466	1.4124	0.2364	2.5884	471	467	480	99%	124.79	6
65	a b c d ac bc bd cd	16	3	73	466	2.7492	1.2083	4.2901	482	469	539	97%	129.63	6
66	a b c d ad bc bd cd	17	4	58	466	2.7887	1.5248	4.0527	483	470	524	96%	104.64	6
67	a b c d ab ac ad bc bd	21	11	36	466	3.0308	2.4785	3.5831	487	477	502	96%	25.95	5
68	a b c d ab ac ad bc cd	34	6	161	466	3.5116	1.9423	5.0808	500	472	627	93%	37.21	5
69	a b c d ab ac ad bd cd	6	1	17	466	1.6462	0.4634	2.8290	472	467	483	99%	48.91	5
70	a b c d ab ac bc bd cd	29	5	134	466	3.3398	1.7857	4.8939	495	471	600	94%	38.11	5
71	a b c d ab ad bc bd cd	45	10	195	466	3.8061	2.3420	5.2701	511	476	661	91%	25.29	5
72	a b c d ac ad bc bd cd	32	6	146	466	3.4589	1.9353	4.9824	498	472	612	94%	101.28	5
73	a b c d ab ac ad bc bd cd	41	8	185	466	3.7086	2.1969	5.2203	507	474	651	92%	25.13	4
74	a b c d ab ac ad bc bd cd abc	355	32	3866	466	5.8719	3.4841	8.2598	821	498	4332	57%	8.63	3
75	a b c d ab ac ad bc bd cd abd	25	5	110	466	3.1800	1.6600	4.7000	491	471	576	95%	18.54	3
76	a b c d ab ac ad bc bd cd acd	1	0	1	466	-16.9451	-19.3220	-14.5683	467	466	467	100%	19.39	3
77	a b c d ab ac ad bc bd cd bcd	1	0	1	466	-16.7498	-19.1004	-14.3991	467	466	467	100%	20.42	3
78	a b c d ab ac ad bc bd cd abc abd	469	42	5219	466	6.1489	3.7378	8.5600	935	508	5685	50%	0.00	2
79	a b c d ab ac ad bc bd cd abc acd	39	3	425	466	3.6624	1.2746	6.0503	505	469	891	92%	8.63	2
80	a b c d ab ac ad bc bd cd abc bcd	55	5	596	466	4.0011	1.6133	6.3890	521	471	1062	89%	8.63	2
81	a b c d ab ac ad bc bd cd abd acd	1	0	1	466	-17.2484	-19.7015	-14.7953	467	466	467	100%	14.01	2
82	a b c d ab ac ad bc bd cd abd bcd	1	0	1	466	-18.6918	-21.0855	-16.2980	467	466	467	100%	8.70	2
83	a b c d ab ac ad bc bd cd acd bcd	0	0	0	466	-40.1284	-43.0059	-37.2509	466	466	466	100%	12.54	2
84	a b c d ab ac ad bc bd cd abc abd acd	33	2	359	466	3.4708	1.0597	5.8819	499	468	825	93%	0.00	1
85	a b c d ab ac ad bc bd cd abc abd bcd	10	0	107	466	2.2573	-0.1538	4.6684	476	466	573	98%	0.00	1
86	a b c d ab ac ad bc bd cd abc acd bcd	6	0	66	466	1.7916	-0.5963	4.1794	472	466	532	99%	8.63	1
87	a b c d ab ac ad bc bd cd abd acd bcd	0	0	0	466	-41.4525	-44.1925	-38.7125	466	466	466	100%	2.62	1
88	a b c d ab ac ad bc bd cd abc abd acd bcd	1	0	8	466	-0.4210	-2.8321	1.9901	467	466	474	100%	0.00	0
89	a b c d ab ad bc bd abc	22	12	37	466	3.0681	2.5336	3.6026	488	478	503	95%	18.47	5
90	a b c d ab ad bc bd abd	21	11	35	466	2.9968	2.4602	3.5333	487	477	501	96%	19.01	5
91	a b c d ab ad bc bd acd	23	13	39	466	3.1061	2.5705	3.6418	489	479	505	95%	25.11	5
92	a b c d ab ad bc bd abc abd	21	11	35	466	2.9968	2.4602	3.5333	487	477	501	96%	11.20	4

a = Cochrane, b = PLUS, c = Medline, d = Renal

Nephrology Study: Estimating the Horizon
Comparison of Various Models

Model#	Model evaluated	Missing cell Estimate	Missing cell Lower 95% CI	Missing cell Upper 95% CI	# of known journals	Coefficient estimate	Coefficient Lower 95% CI	Coefficient Upper 95% CI	Horizon estimate	Horizon Lower 95% CI	Horizon Upper 95% CI	Known journals as % of Horizon	Deviance	DF
93	a b c d ab ad bc bd abc acd	24	13	40	466	3.1358	2.6020	3.6696	490	479	506	95%	10.03	4
94	a b c d ab ad bc bd abd acd	21	11	35	466	3.0115	2.4737	3.5492	487	477	501	96%	18.49	4
95	a b c d ab ad bc bd abc abd acd	21	12	36	466	3.0261	2.4886	3.5637	487	478	502	96%	2.57	3

a = Cochrane, b = PLUS, c = Medline, d = Renal

DISCUSSION

A researcher typically conducts a systematic review of the literature in an attempt to identify as many relevant journals or articles upon which to base a future research endeavor. Such a review would identify a number of journals, but it probably would not identify all journals in the area of interest. Knowing this, the researcher uses the information collected to estimate the number of journals that might have been missed. If that number is small relative to the number of journals identified in the review she would be confident in knowing that she has the most current information to base her next study upon. If the number is large relative to the number of journals she identified, she might want to allocate more resources to the review.

The results of the example presented here suggest that the investigator identified about 99% of the Horizon journals following a review of three databases, and 96% following a review of four databases. These are sufficiently high percentages to convince the authors that the review identified the majority of the targeted journals and that searching additional databases would probably be unnecessary. But these findings illustrate another point.

You can see from the summary in TABLE 5 that following the review of the first two databases, 251 known journals had been identified and the Horizon estimate was 268. When the third database is added to the review, the number of known journals and the Horizon estimate jump to 448 and 449 respectively. Finally, after completing the search of the fourth database, 466 known journals had been identified and the Horizon estimate was 487. Clearly the estimate of the Horizon following the completion of the 2 and 3 database reviews are underestimates of the Horizon since those estimates are less than the number of known journals following the review of all 4 databases.

So, how do you know when you are converging on the Horizon? You will never know with certainty, but as each additional database is searched the number of new hits should diminish. This should be the case if the researcher was fortunate enough to search the databases in the order of descending number of hits. Thus, when the number of additional journals found for each additional database searched diminishes to a small percentage, say to less than 10% of the total number of known journals, and the total number of hits is a large percentage of the Horizon estimate (e.g., greater than 90%), you can be confident that the Horizon is not far off. In the current example, the fourth search identified 18 previously unidentified journals, which is less than 4% of the total number of known journals. In combination with the knowledge that the total number of hits was about 96% of the Horizon we feel that the investigator has sufficient evidence to stop her review after searching four databases. Had the first percentage exceeded 10% or the second been smaller than 90%, we would recommend adding a fifth database to the review.

Could these 4 databases have been searched in a more efficient order?
The database search order can impact the speed with which the review converges on the Horizon. For a review involving four databases, 4 factorial, or 24 different search orders are possible. If resources are limited, it would be extremely helpful to know which database is likely to produce the most hits, and to search it first. In the current example it would have been more efficient to search Medline first as 444 (95%) of the 466 known journals were found there. Had the investigator searched Renal next she would have identified 463 journals, which is 97% of the Horizon estimate. When search results from Plus are added, all 466 journals would have been identified and the best model arising from searching these 3 databases provides a Horizon estimate of 489 with 95% CI equal to [479, 505]. The additional 3 hits identified by the third search represents less than 1% of all hits. Thus, with both conditions being satisfied the investigator could stop her review after completing the search of three databases. The results arrived at by using the data from 3 of the 4 searches are almost identical to those of the 4 database review considered above. Adding the Cochrane search as a fourth database to this review would identify no new journals and would be of no help. Of course there is no way of knowing which database will provide the most hits in advance of conducting a review. But knowing now what we have learned from this review, when starting a new review it would probably be a more efficient strategy to search Medline first and then Renal for a nephrology topic.

CONCLUSION

When a researcher embarks on a research study she typically begins with a thorough review of what is known in the area. Such a review often includes a systematic review of the literature but it rarely includes an estimate of the number of articles or journals missed by such a review.

The method outlined in this paper reveals one way to estimate the number of journals or articles a researcher has missed in a systematic review of the literature, and from that obtain an estimate of the Horizon. With these estimates the researcher would have a basis upon which to decide if the review was sufficient or if further searching is recommended. Also, the importance of search order has been discussed and it has been shown that some orders are more efficient than others. Indeed, if the actual search order was equivalent to the optimal search order, the search of one or more databases may have been unnecessary, resulting in the demand of fewer resources.

In its current state, the macro could be used to generate Horizon estimates for reviews involving fewer than five databases, but it would need to be modified to accommodate reviews involving five or more databases.

REFERENCES

Spoor P, Airey M., Bennett C., Greenhill J., and R. Williams. 1996. Use of capture-recapture technique to evaluate the completeness of systematic literature searches. BMJ 313:342-343.

ACKNOWLEDGMENTS

Many thanks to Haynes RB, Garg AX, McKibbon KA, Wilczynski NL, Kastner M, Iansavichus AV, and Walters LA, for providing the nephrology database for this analysis.

CONTACT INFORMATION

Gary Foster PhD
St. Joseph's Healthcare Hamilton
50 Charlton Avenue East, Room H323
Hamilton ON Canada
L8N 4A6

E-mail: fosterg@mcmaster.ca