Supplemental data file

Inherent limitations of rare event detection [1-3]

Three principle issues in rare event detection were evaluated. First, what is the lowest number of CTC that would need to be in a tube of blood to detect one CTC? Second, what is the theoretical level of variability in measuring the reproducibility of rare events based on a random distribution? Third, if one obtains a given result in clinical practice, what is the range of CTC numbers that might have actually been in the tube based on statistical considerations?

The probability of a random sample of size n containing x events in a total of n tests using the *binomial distribution* is given by the following formula:

$$P(x) = \frac{n!}{x!(n-x!)} p^x(1-p)^{n-x} = \binom{n}{x} p^x(1-p)^{n-x}$$

Where: P(x) = the probability of an event (x) in a unit of space

x = number of events

p = probability of detecting (or observing) x events

$1-p$ = probability of not detecting (or observing) x events

n = sample size

The mean μ and variance σ^2 of the binomial distribution are given by the following:

$$\mu = np \quad \text{and} \quad \sigma^2 = np(1-p)$$

The minimum average number of CTCs (n) required to be present in a single 7.5mL sample of blood to ensure the detection on average of at least 1 CTC (μ) given an average assay recovery of CTCs spiked into 7.5mL of blood of ~85% (p), is therefore:

$$\mu = np$$

$$1 = n(85\%)$$

$$n = 1 / 0.85 \approx 1.2 \text{ CTC}$$

The standard deviation σ for this value of *n* would be determined as follows:

$$\sigma = \sqrt{np(1-p)} = \sqrt{1.2 * 0.85(1-0.85)} = 0.4$$

These results indicate that in order to detect on average 1 CTC with an average assay recovery of 85%, a 7.5mL blood sample would have to contain on average 1.2 CTCs.

For CTC detection, imagine a volume of blood that has been divided into CTC size units. This creates a very large sample size *n*, with a very small probability p of any single volume *n* containing an event *x* (i.e. a CTC). In this situation, with a large *n* and a small p, the *Poisson distribution* can be used to approximate the binomial probability. The *Poisson distribution* is important in describing random (or rare) occurrences where each sample (or volume) *n* has an equal probability of containing an event *x*, such as is the case with the distribution of CTC in a volume of blood.

The probability of a random sample of size *n* containing *x* events can be calculated using the *Poisson distribution* and is given by the following formula:

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

An interesting and useful property of a *Poisson distribution* is that the variance σ^2 is equal to the mean μ. This would make the standard deviation equal to $\sqrt{\mu}$, and the theoretical coefficient of variation (%CV) equal to $\frac{\sqrt{\mu}}{\mu}$.

Using the above %CV formula, at CTC counts of 4, 18, 71, 286, and 1142, the inherent %CVs of actually counting those numbers of events would be predicted to be 50%, 24%, 12%, 6%, and 3%, respectively. These predicted %CVs are very similar to the observed %CVs of 47%, 22%, 11%, 2%, and 5%, respectively, shown in **Table 1**. These findings suggest that the CellSearch assay does not add additional variation to the inherent variation of counting random events due to the *Poisson distribution*.

When calculating confidence intervals (CI) for rare events (i.e. CTC counts), one must keep in mind that the *Poisson distribution* assumes the shape of a *normal distribution* when the number of events is greater than about 100. So we use a *Poisson distribution* for rare events (when the number of events is less than 100), but when the number of events is greater than 100, we can use a modified formula from the *normal distribution* to determine the 95% CI's.

Table 2 provides the lower and upper *confidence factors* used to calculate an exact 95% CI based on a specified number of events (or counts), from 1 to 100. To calculate the exact 95% CI, multiply the number of events (or counts) by the associated *confidence factors* and add these values separately to the count. For example, in **Table 1**, the average observed number of CTC at the 18 CTC spike was 22 CTC (122% recovery). The lower and upper confidence limits are calculated using the *confidence factors* provided in **Table 2**. The factors for 22 events are 0.6267 and 1.5140 for the lower and upper limits of the 95% CI, respectively. Therefore, the exact 95% CI for the average CTC count of 22 would be:

Lower limit = 22(0.6267) = 13.8

Upper limit = 22(1.5140) = 33.3

Thus, for the average % recovery of 122% (22 / 18 CTC)

Lower limit = (14 / 18) * 100% = 77.7%

Upper limit = (33 / 18) * 100% = 183.3%

95% C.I. for average of 122% recovery = 78% to 183%

The formula for the calculation of an approximate 95% CI for a *Poisson distribution* with more than 100 counts μ is:

$$\text{Approximate C.I.} = \mu \pm z_\alpha \sqrt{\mu}$$

Where: z_α = 1.645 for a 90% CI, 1.96 for a 95% CI, or 2.58 for a 99% CI

Lastly, similar considerations apply to the issue of estimating the range of CTC numbers when a given number is measured by the assay. Recall that for a *Poisson distribution* the variance σ^2 is equal to the mean μ, which would make the standard deviation equal to the square root of the mean $\sigma = \sqrt{\mu}$. For a sample size of *n*=1, σ is indeterminate, as we have no knowledge of σ from a single determination (x_1). Although σ is unknown, it is possible to determine the true mean value μ within a certain confidence interval $[\mu_1, \mu_2]$. For $n \to \infty$, a *Poisson distribution* with a mean μ and standard deviation σ is known. If we take one sample from this distribution (*n*=1), this sample will contain x_1 number of CTCs. If we assume that this sample falls within a given confidence interval (z_α), the true average falls within $[\mu_1, \mu_2]$ with the same given confidence, if μ_1 and μ_2 are defined as follows:

$$x_1 = \mu_1 - z_\alpha \sqrt{\mu_1} \text{ and } x_1 = \mu_2 + z_\alpha \sqrt{\mu_2}$$

when you solve the above equation for μ_1 and μ_2, you get

$$\mu_1 = (x_1 + z_\alpha) - \frac{\sqrt{(2z_\alpha + 2x_1)^2 - 4x_1{}^2}}{2}$$

$$\mu_2 = (x_1 + z_\alpha) + \frac{\sqrt{(2z_\alpha + 2x_1)^2 - 4x_1{}^2}}{2}$$

Figure 1 shows μ_1 and μ_2 for a 95% CI (z_α = 1.96, solid line), the 68% CI (z_α = 1.00, short dashed line), and the 38% CI (z_α = 0.50, long dashed line) for x_1 values of 0 to 25 CTC. The range of the true average, μ, based on a single blood draw resulting in x_1 number of CTC, can be read from **Figure 1** with 38%, 68%, and 95% confidence. For example when 5 CTC are detected (x_1=5), you can be 95% confident that the true average lies between the 2 and 12 CTC, 68% confident that the true average lies between 3 and 9 CTC, and 38% confident that the true average lies between 3 and 8 CTC.

REFERENCES

1. Motulsky, H. Intuitive Biostatistics, pp. 245-249. New York: Oxford University Press, 1995.
2. Box, G.E.P., Hunter, W.G., and Hunter, J.S. Statistics for experimenters. An introduction to design, data analysis, and model building, pp. 137-145. New York: John Wiley and Sons, 1978.
3. Daly, L. Simple SAS macros for the calculation of exact binomial and Poisson confidence limits. Comput. Biol. Med., *22:* 351-361, 1992.

Table 1. Method accuracy measured by recovery of SKBR-3 tumor cells spiked into 7.5 mL blood of 5 healthy donors

Expected CTC Count	Observed CTC Count			% Recovery		%CV
	Average	StDev	95% C.I.	Average	95% C.I.	
4	4	2	1 - 11	110	25 - 275	47
18	22	5	14 - 33	122	78 - 183	22
71	70	8	55 - 88	99	77 - 124	11
286	247	5	216 - 277	86	76 - 97	2
1142	971	46	910 - 1032	85	80 - 90	5

Table 2. 95% Confidence Interval Factors for Poisson-Distributed Events

number of events	95% CI, Lower Limit Factor	95% CI, Upper Limit Factor		number of events	95% CI, Lower Limit Factor	95% CI, Upper Limit Factor
0	0.0000	3.7000		51	0.7446	1.3148
1	0.0253	5.5716		52	0.7468	1.3114
2	0.1211	3.6123		53	0.7491	1.3080
3	0.2062	2.9224		54	0.7512	1.3048
4	0.2725	2.5604		55	0.7533	1.3016
5	0.3247	2.3337		56	0.7554	1.2986
6	0.3670	2.1766		57	0.7574	1.2956
7	0.4021	2.0604		58	0.7593	1.2927
8	0.4317	1.9704		59	0.7612	1.2899
9	0.4573	1.8983		60	0.7631	1.2872
10	0.4795	1.8390		61	0.7649	1.2845
11	0.4992	1.7893		62	0.7667	1.2820
12	0.5167	1.7468		63	0.7684	1.2794
13	0.5325	1.7100		64	0.7701	1.2770
14	0.5467	1.6778		65	0.7718	1.2746
15	0.5597	1.6493		66	0.7734	1.2722
16	0.5716	1.6239		67	0.7750	1.2700
17	0.5825	1.6011		68	0.7765	1.2677
18	0.5927	1.5804		69	0.7781	1.2656
19	0.6021	1.5616		70	0.7795	1.2634
20	0.6108	1.5444		71	0.7810	1.2614
21	0.6190	1.5286		72	0.7824	1.2593
22	0.6267	1.5140		73	0.7838	1.2573
23	0.6339	1.5005		74	0.7852	1.2554
24	0.6407	1.4879		75	0.7866	1.2535
25	0.6471	1.4762		76	0.7879	1.2516
26	0.6532	1.4652		77	0.7892	1.2498
27	0.6590	1.4549		78	0.7905	1.2480
28	0.6645	1.4453		79	0.7917	1.2463
29	0.6697	1.4362		80	0.7929	1.2446
30	0.6747	1.4276		81	0.7941	1.2429
31	0.6795	1.4194		82	0.7953	1.2413
32	0.6840	1.4117		83	0.7965	1.2397
33	0.6884	1.4044		84	0.7976	1.2381

Table 2 (con't). 95% Confidence Interval Factors for Poisson-Distributed Events

number of events	95% CI, Lower Limit Factor	95% CI, Upper Limit Factor		number of events	95% CI, Lower Limit Factor	95% CI, Upper Limit Factor
34	0.6925	1.3974		85	0.7988	1.2365
35	0.6965	1.3908		86	0.7999	1.2350
36	0.7004	1.3844		87	0.8010	1.2335
37	0.7041	1.3784		88	0.8020	1.2320
38	0.7077	1.3726		89	0.8031	1.2306
39	0.7111	1.3670		90	0.8041	1.2292
40	0.7144	1.3617		91	0.8051	1.2278
41	0.7176	1.3566		92	0.8061	1.2264
42	0.7207	1.3517		93	0.8071	1.2251
43	0.7237	1.3470		94	0.8081	1.2237
44	0.7266	1.3425		95	0.8091	1.2224
45	0.7294	1.3381		96	0.8100	1.2212
46	0.7321	1.3339		97	0.8109	1.2199
47	0.7348	1.3298		98	0.8118	1.2187
48	0.7373	1.3259		99	0.8128	1.2175
49	0.7398	1.3221		100	0.8136	1.2163
50	0.7422	1.3184				

Appendix Figure 1