Size distribution of clinically detected CRC

When a preclinical cancer initiates, the size of the cancer upon clinical detection (in the absence of screening—ie, the expiration of sojourn time) is determined. CRC-SPIN's size at clinical detection is based on the overall SEER distribution of CRC size from 1975-1979,[1] but the parameterization of this size is not explained. Here, we briefly describe the steps we took to derive this distribution for CRC-AIM.

We conducted a SEER query of the 1975-1979 registry data using the conditions described below:

SEER query for 1975-1979 CRC size

Software

Surveillance Research Program, National Cancer Institute SEER*Stat software (www.seer.cancer.gov/seerstat) version 8.3.5. accessed 07/12/2018

Data

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence – SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2017 Sub (1973-2015 varying) – Linked to County Attributes – Total U.S., 1969-2016 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission.

Selection

Select Only: Malignant Behavior, Known Age, Cases in Research Database

{Site and Morphology Site recode ICD-O-3/WHO 2008}='Colon and Rectum'

AND {Site and Morphology Histologic Type ICD-O-3}=8000-8001,8010,8020,8140,8210-8211,8220-8221,8260-8263,8480-8482,8490

AND {Race, Sex, Year Dx, Registry, County.Year of diagnosis}='1975','1976','1977','1978','1979'

Table

Expanded EOD(1) - CP53 (1973-1982)

Expanded EOD(2) - CP54 (1973-1982)

Expanded EOD(3) - CP55 (1973-1982)

Expanded EOD(4) - CP56 (1973-1982)
Expanded EOD(5) - CP57 (1973-1982)
Expanded EOD(6) - CP58 (1973-1982)
Expanded EOD(7) - CP59 (1973-1982)
Expanded EOD(8) – CP60 (1973-1982)
Expanded EOD(9) – CP61 (1973-1982)
Expanded EOD(10) – CP62 (1973-1982)
Expanded EOD(11) – CP63 (1973-1982)
Expanded EOD(12) – CP64 (1973-1982)
Expanded EOD(13) – CP65 (1973-1982)
SEER historic stage A
2-Digit NS EOD part 1 (1973-1982)
AJCC 5th Ed Schrag Code 1975-1979

A total of 50,743 CRCs were queried. The SEER Extent of Disease (EOD) coding scheme records CRC sizing information in the unit of millimeter: a value of 0 to 9 is recorded in EOD(1) – CP53 (1973-1982) for the value in the tens place, and a value of 0 to 9 is recorded in EOD(2) – CP54 (1973-1982) for the value in the ones place. Although this theoretically allows for CRC sizes up to 99 mm, this is not how the information is represented. Instead, size is actually coded up to 97 mm, with tumors that are greater than or equal to 98 mm coded as “98”.

Additionally, the following special codes are used[2]:

- 00: No mass
- 0&: Microscopic focus or foci only
- -- Not stated

The following criteria were used to filter out results from further analysis: Unstaged CRC (7,692 records) and CRC size where size was recorded as --, 00, 0&,

Blank(s)Blank(s) (14,912 records). Ultimately, a total of 17,258 results were excluded, resulting in 33,485 results (50,743 minus 17,258). Of those 33,485 results, 32,032 CRCs range from 1 mm to 97 mm and 1,453 CRCs are recorded as 98 mm, corresponding to the “≥98 mm” category (**Figure 1**). We treated the 1,453 “≥98 mm” records as missing observations and modeled these values, extrapolating a non-truncated right-tail of the distribution. Specifically, we parametrically modeled the CRC counts from 50 mm to 97 mm and extrapolated the counts modeling past 97 mm until the extrapolated total equaled ~1,453 observations (actual n = 1,484). The extrapolated counts are combined with the original counts and the entire distribution was fit to obtain the probability density function of CRC size.

We plotted the counts of discrete CRC size categories from the SEER registry data and observed that most CRCs in this subsample were rounded to the nearest centimeter (eg, 50 mm, 60 mm, 70 mm, etc.) (**Figure 2**). Another set of CRCs was rounded to the nearest half-centimeter (eg, 55 mm, 65 mm, 75 mm, etc.). Finally, a third set was rounded to the nearest millimeter. Notably, counts rounded to the nearest centimeter are biased because they are inclusive of CRCs rounded to the nearest half-centimeter (eg, a 52 mm CRC rounds to 50 mm) and those rounded to the nearest 1 mm (eg, a 50.4 mm CRC rounds to 50mm). Similarly, the counts on the half-centimeter (eg, 55 mm, 65 mm, etc.) are biased since they are inclusive of counts rounded to the nearest millimeter. These biases were ignored for this simple modeling exercise.

To perform the model extrapolation, we created three separate Poisson regression models for each rounding scenario (**Figure 3**). The Poisson regressions were applied to each missing size group associated with each rounding scenario past

97 mm. For example, the regression for the nearest centimeter rounding scenarios was applied to sizes of 100 mm, 110 mm, etc. Size was increased by millimeter increments until ~1,453 observations were obtained (**Table 1**). Finally, the 32,032 values coded from 1 mm to 97 mm were combined with the extrapolated 1,484 values from 98 mm to 140 mm (**Figure 4**). The probability density function (PDF) of the generalized log distribution is sampled to generate a CRC size at clinical diagnosis from 1 mm to 140 mm.

Figure 1. Histogram of clinically diagnosed colorectal cancer (CRC) sizes from 1975-1979 SEER database. Tumors within the “≥98 mm” category are represented by the patterned bar.

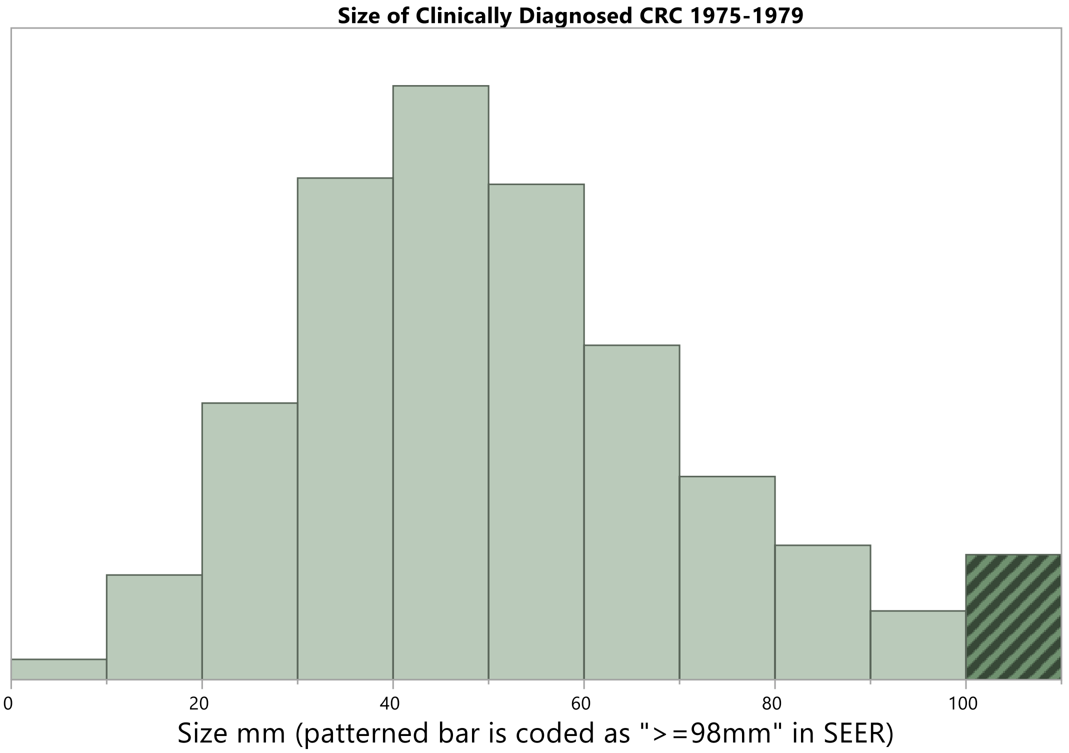


Figure 2. Count of clinically diagnosed colorectal cancer (CRC) sizes from 1975-1979 SEER database. Only CRCs between 50 mm and 97 mm are included.

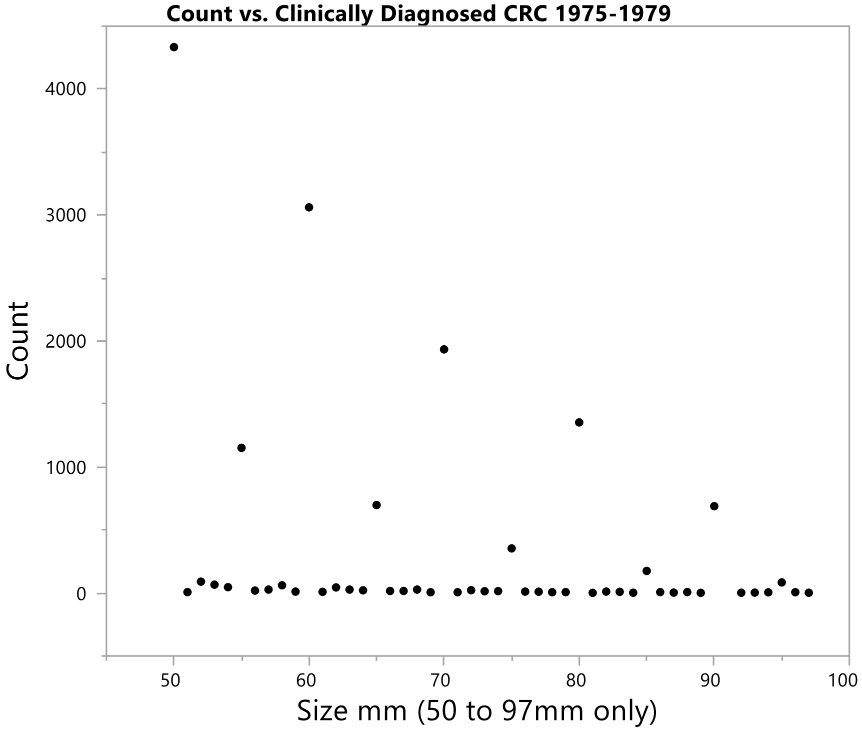
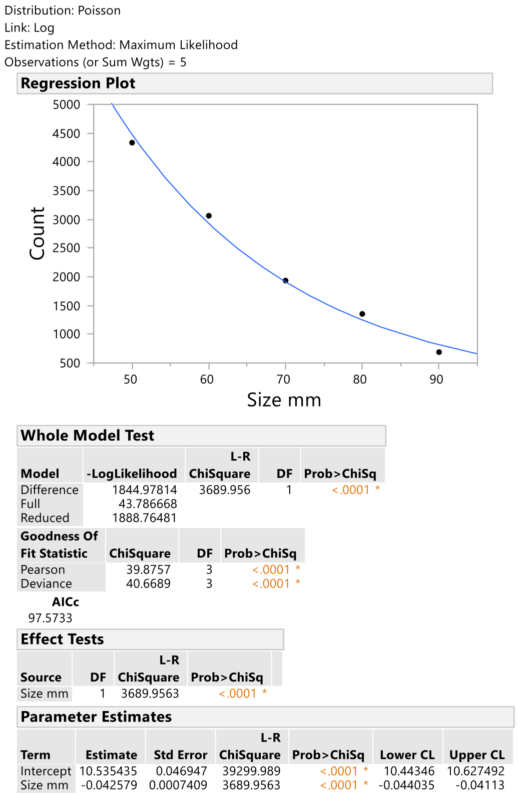


Figure 3. Poisson regression models for colorectal cancer (CRC) size distribution rounding scenarios. Models extrapolating sizes based on (A) rounding to the nearest centimeter (eg, 50 mm, 60 mm, 70 mm, etc.); (B) rounding to the nearest half-centimeter (eg, 55 mm, 65 mm, 75 mm, etc.), excluding scenario (A); and (C) rounding to the nearest millimeter, excluding scenarios (A) and (B).

(A)



Distribution: Poisson
Link: Log
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 5

Regression Plot

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	1844.97814	3689.956	1	<.0001 *
Full	43.786668			
Reduced	1888.76481			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq
Pearson	39.8757	3	<.0001 *
Deviance	40.6689	3	<.0001 *

AICc
97.5733

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
Size mm	1	3689.9563	<.0001 *

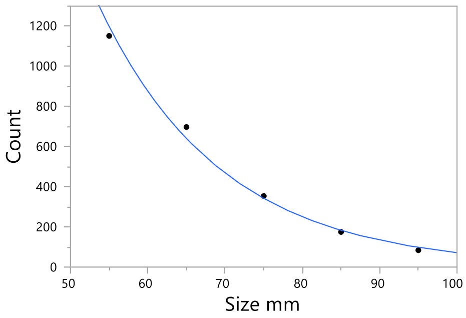
Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq	Lower CL	Upper CL
Intercept	10.535435	0.046947	39299.989	<.0001 *	10.44346	10.627492
Size mm	-0.042579	0.0007409	3689.9563	<.0001 *	-0.044035	-0.04113

(B)

Distribution: Poisson
Link: Log
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 5

Regression Plot



Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	770.490678	1540.981	1	<.0001 *
Full	23.9379646			
Reduced	794.428642			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq
Pearson	9.6399	3	0.0219 *
Deviance	9.6204	3	0.0221 *

AICc
57.8759

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
Size mm	1	1540.9814	<.0001 *

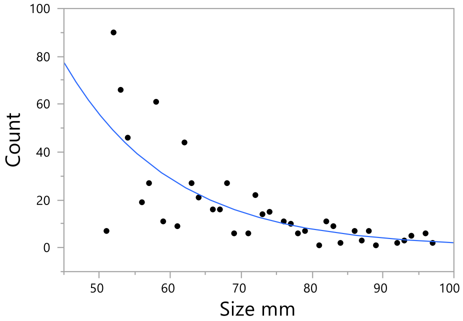
Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq	Lower CL	Upper CL
Intercept	10.533062	0.1163475	7923.9396	<.0001 *	10.305856	10.761979
Size mm	-0.062621	0.0017848	1540.9814	<.0001 *	-0.066146	-0.059149

(C)

Distribution: Poisson
Link: Log
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 37

Regression Plot



Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	203.527025	407.0540	1	<.0001 *
Full	191.637037			
Reduced	395.164061			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq
Pearson	205.9491	35	<.0001 *
Deviance	229.1494	35	<.0001 *

AICc
387.6270

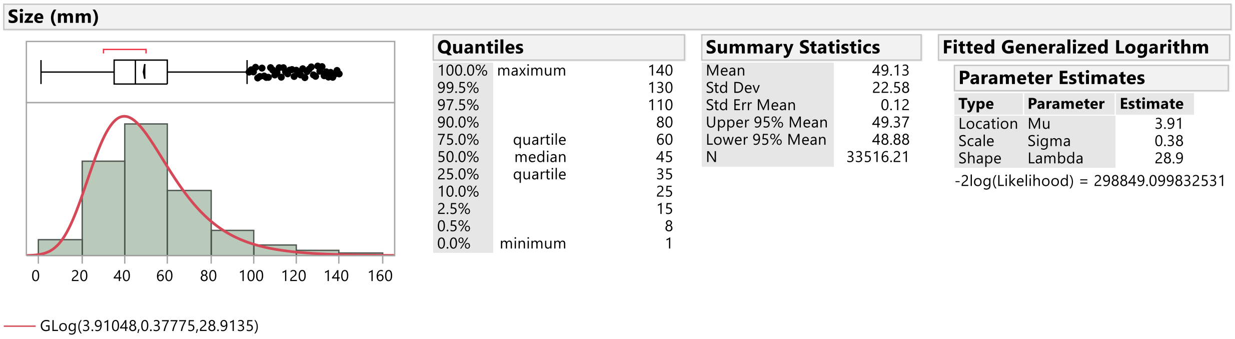
Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
Size mm	1	407.05405	<.0001 *

Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq	Lower CL	Upper CL
Intercept	7.3115577	0.231883	1036.0574	<.0001 *	6.8598575	7.7691307
Size mm	-0.065812	0.0036326	407.05405	<.0001 *	-0.07303	-0.058786

Figure 4. Final model size distribution of colorectal cancer (CRC) at clinical detection for CRC-AIM.



Quantiles

100.0%	maximum	140
99.5%		130
97.5%		110
90.0%		80
75.0%	quartile	60
50.0%	median	45
25.0%	quartile	35
10.0%		25
2.5%		15
0.5%		8
0.0%	minimum	1

Summary Statistics

Mean	49.13
Std Dev	22.58
Std Err Mean	0.12
Upper 95% Mean	49.37
Lower 95% Mean	48.88
N	33516.21

Fitted Generalized Logarithm

Parameter Estimates

Type	Parameter	Estimate
Location	Mu	3.91
Scale	Sigma	0.38
Shape	Lambda	28.9

-2log(Likelihood) = 298849.099832531

Table 1. Extrapolated counts of colorectal cancer size from three Poisson models for different rounding scenarios. The number of size buckets was increased until ~1,453 overall observations were obtained (actual n = 1,484). Values rounded for visual simplicity. n/a = not applicable

Size (mm)	Counts from Poisson Model 1 (nearest centimeter)	Counts from Poisson Model 2 (nearest half-centimeter)	Counts from Poisson Model 3 (nearest millimeter)
98	n/a	n/a	2
99	n/a	n/a	2
100	532	n/a	n/a
101	n/a	n/a	2
102	n/a	n/a	2
103	n/a	n/a	2
104	n/a	n/a	2
105	n/a	52	n/a
106	n/a	n/a	1
107	n/a	n/a	1
108	n/a	n/a	1
109	n/a	n/a	1
110	348	n/a	n/a
111	n/a	n/a	1
112	n/a	n/a	1
113	n/a	n/a	1
114	n/a	n/a	1
115	n/a	28	n/a
116	n/a	n/a	1
117	n/a	n/a	1
118	n/a	n/a	1
119	n/a	n/a	1
120	227	n/a	n/a
121	n/a	n/a	1
122	n/a	n/a	0
123	n/a	n/a	0
124	n/a	n/a	0
125	n/a	15	n/a
126	n/a	n/a	0
127	n/a	n/a	0

128	n/a	n/a	0
129	n/a	n/a	0
130	148	n/a	n/a
131	n/a	n/a	0
132	n/a	n/a	0
133	n/a	n/a	0
134	n/a	n/a	0
135	n/a	8	n/a
136	n/a	n/a	0
137	n/a	n/a	0
138	n/a	n/a	0
139	n/a	n/a	0
140	97	n/a	n/a

References

1. CISNET Colorectal Cancer Collaborators. RAND Corporation (CRC-SPIN), 2015. HI.001.03112015.70373:https://cisnet.cancer.gov/colorectal/profiles.html.

2. Ryan RF, Axtell LM, Green SB, et al. *Extent of Disease: Codes and Coding Instructions for the Cancer Surveillance, Epidemiology and End Results Reporting (SEER) Program.* U.S. Department of Health, Education, and Welfare; 1977.