"Moreover, if all these things were clearly, and truly known (which I have but guessed at) it would appear, how small a part of the People work upon necessary Labours, and Callings, viz. how many Women, and Children do just nothing, onely learning to spend what others get? how many … live by puzling poor people with unintelligible Notions in Divinity, and Philosophie, how many by perswading credulous, delicate, and Litigious Persons, that their Bodies, or Estates are out of Tune, and in danger. And on the other side, how few are employed in raising, and working necessary food and covering? and of the speculative men, how few do truly studie *Nature*, and *Things*? The more ingenious not advancing much further than to write, and speak wittily about these matters.

— John Graunt, *Natural and Political Observations, London, 1662*

Death in London: Establishing Credibility — Who, What, Where, Why, When, and How?

More data: Figure 1 (facing) presents data on the causes of death in London, in 1632. That's the target of my next data analysis and what I want to extract from the facts is information: I want to "know" what people died of in London three to four hundred years ago. And now let's take it through the steps: For these data I need orientation, Who, What, Where, … and I need technique — this is more difficult than the data for breakfast cereals.

So, death in London, Who, What, Where: The number one question is probably "Why London, and why 1632?" I chose it because, according to *The World of Mathematics"* where I found these data, the original work

by John Graunt, was one of the original uses of statistics for data analysis[1] It seemed appropriate to begin a discussion of data analysis by presenting one of the true beginnings of the profession. Even such "obvious" techniques as these began somewhere and had to be invented by a human mind like our own, like the mind of John Graunt.

Continuing, before the numbers, what are these data about and what's the context? I've already given you that: I've told you the data are about death and I've warned you that the data are very old, from another time. Still, I want to call attention to the obvious in order to warn you again: These are not numbers, these are data. Don't think about them as numbers, two plus two equals four. Think about life and death and old London — long before the invention of drugs that have kept me, for one, alive, before Pasteur with his understanding of bacteria, and only slightly after the invention of the microscope.

We also need to know what John Graunt was up to, it may have affected his results: Graunt seems to have known that he was an innovator, for whom the tabulation of diseases and casualties was just the beginning. His interest was in the state, England. He seems to have been interested in the efficient and balanced functioning of the state:

> I conclude, That a knowledge of all these particulars, and many more, whereat I have shot but at rovers, is necessary in order to good, certain, and easie Government, and even to balance Parties, and factions both in *Church* and *State.*
>
> — John Graunt, *op. cit*.

O.K., now — the numbers for death in London.

---

1 From "Foundations of Vital Statistics", by John Graunt, reprinted in "The World of Mathematics", Volume 3, page 1421, published by Simon and Schuster, New York, 1956.

Death in London

The Diseases, and Casualties this year being 1632.

Abortive, and Stilborn . .	445	Grief	11
Affrighted	1	Jaundies	43
Aged	628	Jawfaln	8
Ague	43	Impostume	74
Apoplex, and Meagrom	17	Kil'd by several accidents	46
Bit with a mad dog	1	King'sEvil	38
Bleeding	3	Lethargie	2
Bloody flux, scowring, and flux	348	Livergrown	87
		Lunatique	5
Brused, Issues, sores, and ulcers,	28	Made away themselves	15
		Measles	80
Burnt, and Scalded	5	Murthered	7
Burst, and Rupture	9	Over-laid, and staved at nurse	7
Cancer, and Wolf	10		
Canker	1	Palsie	25
Childbed	171	Piles	1
Chrisomes, and Infants	2268	Plague	8
Cold, and Cough	55	Planet	13
Colick, Stone, and Strangury	56	Pleurisie, and Spleen	36
Consumption	1797	Purples, and spotted Feaver	38
Convulsion	241	Quinsie	7
Cut of the Stone	5	Rising of the Lights	98
Dead in the street, and starved	6	Sciatica	1
		Scurvey, and Itch	9
Dropsie, and Swelling	267	Suddenly	62
Drowned	34	Surfet	86
Executed, and prest to death	18	Swine Pox	6
Falling Sickness	7	Teeth	470
Fever	1108	Thrush, and Sore mouth	40
Fistula	13	Tympany	13
Flocks, and small Pox	531	Tissick	34
French Pox	12	Vomiting	34
Gangrene	5	Worms	27
Gout	4		

Christened	Males 4994 Females 4590 In all 9584		Buried	Males 4932 Females 4603 In all 9535	Whereof, of the Plague 8

Increased in the Burials in the 122 Parishes, and at the Pest-house this year	993
Decreased of the Plague in the 122 Parishes, and at the Pest-house this year	266

(From Newman, *The World of Mathematics*, Page 1425)

Death in London:
Stem and Leaf — More Technique

The stem and leaf drawings used for the breakfast cereals required a little bit of experimentation with the stems and a lot of experimentation with the leaves, choosing the form that squeezed the most information out of the data. These new data present a pattern that is very common highly resistant to this stem and leaf technique. With these data no simple expansion or compression of the stems, or simple experimentation with the labels is going to get it really "right". Instead, they require the use of several different scales, combined together in one drawing. As before, let me demonstrate the technique (and the problem) with raw numbers, plucked out of context, in order to work with technique. Here is the new set of numbers, shown in Figure 1.

445, 1, 628, 43, 17, 1, 3, 348, 28, 5, 9, 10, 1, 171, 2268, 55, 56, 1797, 241, 5, 6, 267, 34, 18, 7, 1108, 13, 531, 12, 5, 4, 11, 43, 8, 74, 46, 38, 2, 87, 5, 15, 80, 7, 7, 25, 8, 13, 36, 38, 7,98, 1, 9, 62, 86, 6, 470, 40, 13, 34, 1, 27.

Figure 1
More Numbers, for Practice

Suppose I try a straightforward approach, using 10's for my stems: 0, 10, 20, 30, ..., Figure __.

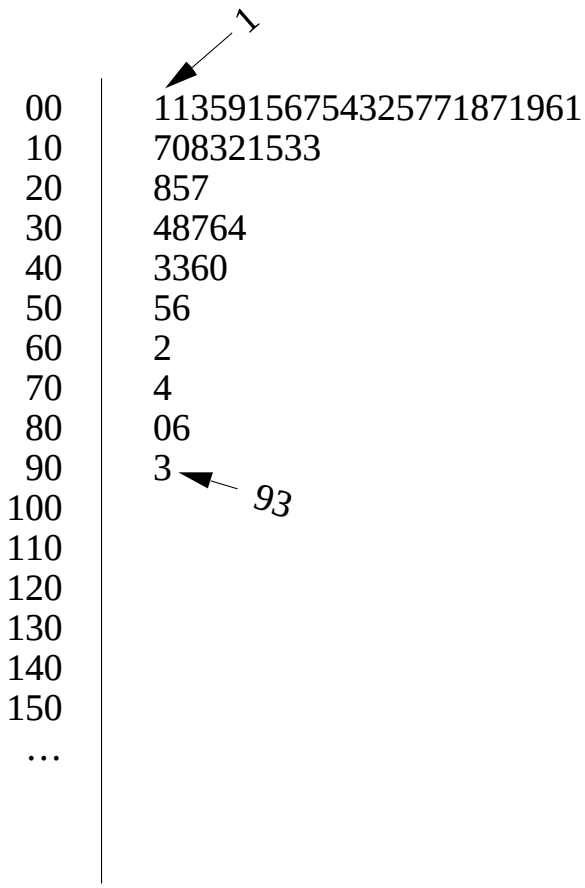


Figure 2
Stem and Leaf, by 10's

No, that's not going to work, I've got to extend the graph all the way up to 2268, which would fill up a couple of pages with this kind of stem.

Well, suppose I try 100's for my stems, compressing the graph, as in Figure 3. That's still going to have too many stems. Worse, it's still too big and it's bunching up a lot of data in a pile, undifferentiated.

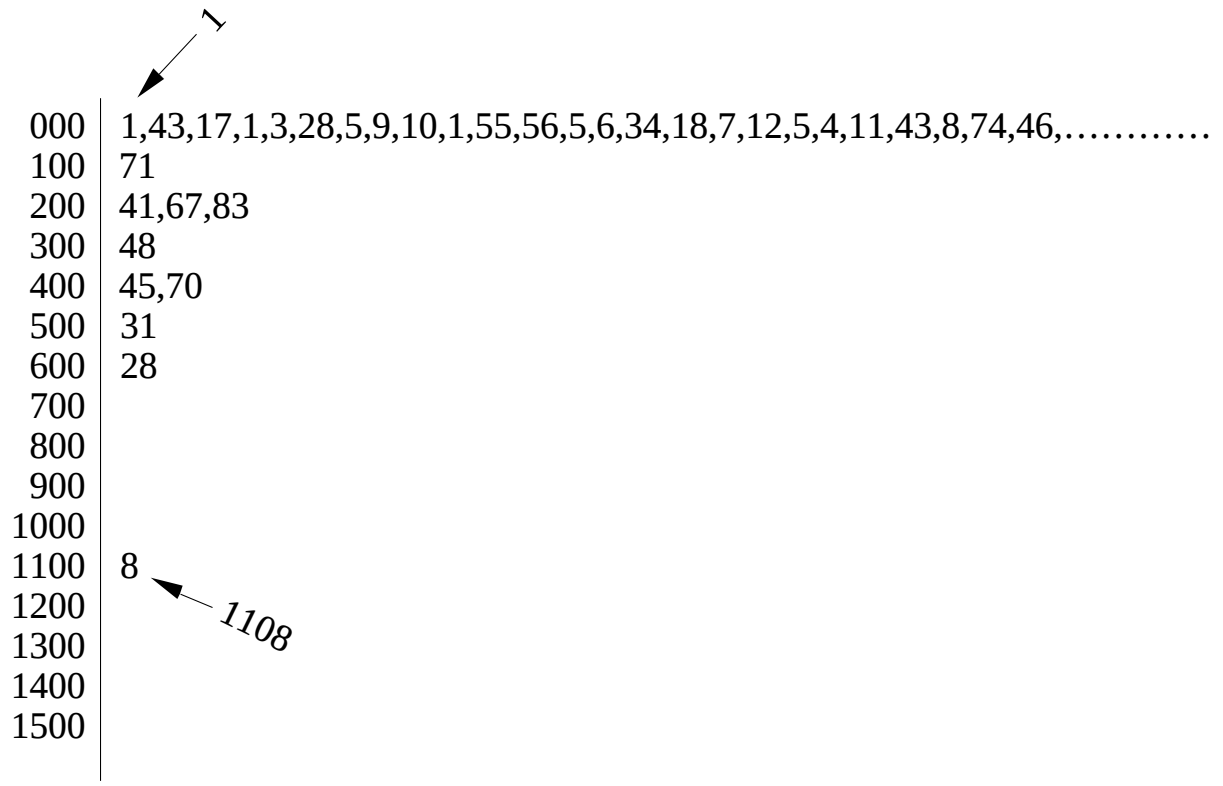


Figure 3
Stem and Leaf, by 100's

Stems of 200 would give me about the right number of stems, but most of the data would be in a pile: In fact, using 200's in Figure 3, and giving each leaf about the same amount of room on the page, I've only been able to squeeze about a quarter of the first set of leaves onto the page. Stems of 500 or 1000 would be worse.

```
   0 |  1, 43, 17,  1,  3, 28,  5,  9,1 0,  1,171, 55, 56,  5,  6,……
 200 |348,241,267,283
 400 |445,531,470
 600 |628
1000 |1108
1200 |
1400 |
1600 |1797
1800 |
2000 |
2200 |2268
2400 |
```

Figure 3
Stem and Leaf, by 200's

There's the problem. And the solution for such a problem is not a great one, but it will do: The solution is to change scale one or more times, right in the middle of the graph. Here in Figure 4, for example, I've counted by 2's up to 10, then by 10's up to 100, then by 100's up to 1000, changing the values.

```
  0 |   1,  1,  1,  1,  1
  2 |   3,  4
  4 |   5,  5,  5,  4,  5
  6 |   6,  7,  7,  7,  7,  6
  8 |   9,  8,  8,  9
 10 |  17, 10, 18, 13, 12, 11, 15, 13, 13
 20 |  28, 34, 38, 25, 36, 38, 34, 27
 40 |  43, 55, 56, 43, 46, 40
 60 |  74, 62
 80 |  80, 98, 86
100 | 171
200 | 348,241,267
400 | 445,531,470
600 | 628
800 |
1000 | 1797,1108
2000 | 2268,2837
```

Figure 4
Stem and Leaf, Variable Stems

Again I'll get a pile up — where I've changed the intervals. And I've actually used seven different scales: By 2's between 0 and 10, by 10's between 10 and 20, by 20's between 20 and 100, by 100's between 100 and 200, by 200's between 200 and 1,000, and by 1,000's between 1,000 and 2,000. That's a rather unpleasant and unsatisfactory solution. And what you'll see later, is that the unpleasant and unsatisfactory feeling you should be getting from these data is itself a clue — to a better solution, later. But for now —

Death In London: The Work

Back to the data — in context. Back to death in London. Now I, for one, am unfamiliar with the diseases listed, including "Rising of the Lights" and "Kings Evil", and there is a tendency to chuckle a little at the peculiarity of these causes of death. I want you to use that chuckle and the peculiarity of the stem and leaf — and every other clue you can muster to make sense of these data: Observe yourself and your reactions to the data. What's going on? Keeping that in mind, let's begin getting a feel for the data.

The first thing I do is simply look at the page. I read some of the labels. Ah, this is unfamiliar turf. I look at the labels on the whole page, "Diseases and Casualties this year being 1632". Ah, mortality statistics filtered through the lens of another culture, before the biology and medicine of our own era. These are not "our" categories, nor our definition of disease. What do I expect from these data? I expect to be mystified, confused by descriptions of disease from a "pre-scientific" era. If anything, I remember something about the plague. Is that relevant? No, only 8 deaths from plague. So I'm expecting confusion, causes as meaningless to me, three hundred years later, as arterial sclerosis would have been to them, three hundred years ago.

Looking at the numbers: They start out with a "445" and continue with a "1" — some of these categories are big, some are small. In fact, they range from over 2,000 down to 1. Looking at the bottom of the page there is some special attention to the plague, but the number dying from it is small. How big is the population experiencing this death. I don't know, but I can guess: The number of people Christened is about ten thousand as is the number buried. If a modern birth or death rate is about 1 or 2 %, then I can guess that London of 1632 already had a

population somewhere between one hundred thousand and one million. That's a rough estimate of the order of magnitude for this population. Let's look at more numbers — and now put some order into the search.

Looking at the detail, I'll use the Stem and Leaf to re-arrange these "diseases" according to size, rare to frequent, and take a look. Even if I had a computer available, one that would "do it right" the first time, I would begin with the stem and leaf, by hand, because it will force me to read labels like "Abortive and Stilborn" and the names of each of these other categories: Doing the stem and leaf, by hand, will rivet my attention on these data. I want about ten stems, I want the definitions of these stems to be easy (so my attention does not get diverted), and so I'll try "stems" beginning at 0, and going up, beginning at 100, and going up, beginning at 200, and going up, ...

0	
100	
200	
300	
400	
500	
600	
700	
800	

. . .

No, that's getting me a lot of categories (as I observed earlier when I treated these numbers as raw numbers). I've got a long way to go before I get to 2,000 (for Chrisomes and Infants). Observing this little numerical difficulty, deciding for myself (without the aid of a computer) how to organize these stems, and finding some difficulty, — I observe my difficulty and learn something about the data: Some of these things are really big, really "off scale" as compared to others. Note this: The "data analysis" lies in observing and learning, not in finishing another "stem and leaf".

So, I'm going to use two scales, perhaps 0 to 1,000 plus a stem for and "Other." So I begin to write

0	Affright, ague, apoplex, bit, bleeding, brused
100	
200	
300	Bloody
400	Abort
500	
600	Aged
700	
800	

. . .

Pausing, with nine leaves, "everything" (six of the first nine) is going in to the first category, failing to differentiate among these things. Let me try again, expanding the stem and leaf.

0	Affright, bit with mad dog, bleeding
10	Apoplex, and Meagrom
20	
30	
40	Ague
50	
60	
70	
80	
. . .	
100	
200	
300	Bloody flux
400	Abort
500	
600	Aged
700	
800	
. . .	
1000	
2000	

O.K., good enough for now. I'll continue using three ranges. And of course the breaks in ranges were chosen for convenience, not because the data break at precisely those points. Continuing:

```
    0 | Affright; bit with mad dog; bleeding; burnt and Scalded; Burst & Rupture;
      | Canker;
   10 | Apoplex, and Meagrom; Cancer and Wolf
   20 | Brused issues
   30 |
   40 | Ague
   50 | Cold and Cough; Colick Stone
   60 |
   70 |
   80 |
      |
. . . _____________________
  100 | Childbed
  200 | Convulsion
  300 | Bloody flux
  400 | Abort
  500 |
  600 | Aged
  700 |
  800 |
      |
. . . _____________________
 1000 | Consumption
 2000 | Chrisomes and Infants;
```

Nineteen leaves — getting messy: The data are piling up in the smallest stem. I'm thinking, "Perhaps I could clean up my technique by introducing yet another range." But for now, I'll just let them pile up and get on with it — I want to look at the data. Beside which, my mind is beginning to focus on these things — which means that the stem and leaf is doing its job: I'm more interested in "Chrisomes", than I am in "bit with mad dog", and I don't want to spend too much time holding myself back with technicalities: I'm trying to learn something: I'm thinking "What are 'chrisomes' ?" and the stem and leaf has me building up a list of names to check as soon as I can find a dictionary that's likely to include such things. In fact, I'm thinking, "I'm not going to go on

automatic and finish this thing, as if I were a computer: I'm going to focus on the big stuff, like Chrisomes. That's where I'll learn something," So, continuing again, but limiting myself to the big stuff:

100	Childbed
200	Convulsion; Dropsie and Swelling
300	Bloody flux
400	Abort; teeth
500	Flocks and small pox
600	Aged
700	
800	
1000	Consumption; Fever
2000	Chrisomes and Infants;

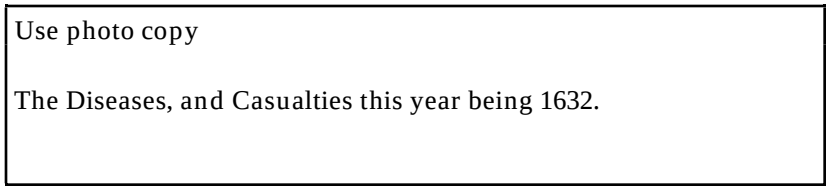
That's simple. Call it "selective" — having ignored most of the data. But that's just being focused. I've decided that the information lies in the big stuff and that's what I'll look at. And what do I see? Well, in reverse order, I still don't know what Chrisomes are, but they have something to do with infants. "Consumption?" Ah, I remember that one: tuberculosis, associated with crowded conditions in the absence of things like clean air and clean water, a public health problem. "Fever?", too general. "Abortive", let's look back for the full label: "Abortive and Stilborn". Ah, we're dealing with childbirth again. "Bloody flux", well— considering that I've already got two categories related to childbirth, I can guess about that one ... and there's "childbed" in a nearby stem. "Teeth"? I'm beginning to guess. I'll bet we're talking about infections, (using a modern definition). And I'm beginning to get a picture: The people at risk are newborns, fertile and child bearing women, and people with some sort of susceptibility to infection: And,

more generally, I'm thinking that this London was a filthy mess in which, if you got an infection, you died.

Now, to the dictionary: I still want to know about "Chrisomes" [ ], O.E.D, — consistent with what I guessed and my hunches about London in 1632: Expose your blood to the environment, expose your mucus membranes to the environment (colick, consumption, pleurisie, sore mouth,) and you die, or you get sick (fever) and then you die.

That is what data analysis is about: Wringing the data for information about the world behind the data, for reasonable hunches and a direction that guides the next step in my study. Right or wrong, so far, it's got me thinking and building hypotheses about this London of 1632.

Death in London: The Report



Tabulation by John Graunt, reproduced from James R. Newman's *The World of Mathematics, Volume 3*, page 1421, published by Simon and Schuster, New York 1956.

Figure 1
Causes of Death in London, 1632

Life in London, three hundred years ago, was dirty and short. That was the picture disclosed in one of the first compilations of vital statistics, published by John Graunt in 1662.

His data described 10,000 deaths from many causes, most of them unknown to the modern vocabulary. But a few causes account for the vast majority of mortality. Most striking, comparing the number of christenings to the number of deaths among the young, infant mortality probably exceeded twenty-five percent (comparing 2,300 deaths from "Chrisomes and Infants" to 9,600 christenings).

Exotic causes that are popular in the current image of historical London accounted for very few of the actual deaths. In this year, for

example, plague accounted for only 8 deaths, By contrast, the major causes of death are appalling familiar to us, even three hundred and fifty years later: Adults die from infections and communicable diseases of which one example, tuberculosis (consumption), accounted for approximately 20% of the mortality. Other categories suggesting infection or contagious disease account for another third of all deaths — including "Fever", "Flocks and Small Pox", "Brused …", "Cold…", "Dropsie …", "Convulsion", "Childbed", "Bloody flux ...", and "Teeth". Adding it up, death related to childbirth, infections, and communicable diseases accounted for approximately eighty percent of mortality.

The data were tabulated by John Graunt from weekly Bills of Mortality and published in 1662. Causes of death were report by attending physicians and bystanders. The report includes detail with respect to sources of data and comments on likely sources of error. Excerpts from Graunt's report are available in reprint as "Foundations of Vital Statistics", in James Newman's *The World of Mathematics*,, noted above. Graunt's stated purpose was to distinguish fact from fiction, to compare true causes of death to those, such as plague, that commanded public attention and for the general purpose of increasing the welfare of the state. Grant's report included extended comments on the quality of the data and likely sources of error.

There's my report, less than five hundred words — probably a bit longer than I would ordinarily write because I have students looking over my shoulder. Now, how did I go from Graunt's table to my summary? The key, and the focus of this discussion of method was the "Stem and Leaf" diagram worked-out earlier. But I haven't even presented it here in my report. It was essential to the process, but once I focused on the few causes of the overwhelming number of deaths, I left it behind. Little would have been added to the reader's knowledge of

death in London if I had recapitulated my steps in the final report. That's hard for a writer: All that work — hidden. No one will know to applaud my diligence. But that' the difference between writing a report about *London* and writing a report about *yourself* — showing your reader how much hard work went into the report. The reader is interested in London. So, spare the reader — and keep the focus.

Exercise: Prepare a report on causes of death in the United States, 1990, using these U.S. Census data.

From the *Statistical Abstract of the United States, 1992*, Table Number 114: Deaths and Death Rates, by Selected Causes, 1990 (1990 estimates are preliminary data from a 10 percent sample including deaths of non-residents.

Cause of Death	**Deaths (1,000), 1990**
All Causes	**2,162.0**
Major cariovascular diseases	920.4
Diseases of heart	725.0
Percent of total	33.5
Rheumatic fever and rheumatic heart disease	6.3
Hypertensive heart disease[1]	23.6
Ischemic heart disease	489.3
Other diseases of endocardium	12.3
All other forms of heart disease	193.5
Hypertension[1]	9.2
Cerebrovascular diseases	145.3
Atherosclerosis	16.5
Other	24.4
Malignancies[2]	506.0
Percent of total	23.4
Of respiratory and intrathoracic organs	143.8
Of digestive organs and peritoneum	121.3
Of genital organs	58.0
Of breast	45.1
Of urinary organs	20.4
Leukemia	18.7
Accidents and adverse effects	93.6
Motor vehicle	47.9
All other	45.7
Chronic obstructive pulmonary diseases and allied conditions	89.0
Bronchitis, chronic and unspecified	3.4
Emphysema	16.5
Asthma	4.6

[1] With or without renal desease
[2] Includes other types of malignancies not shown separately

Other	64.6
Pneumonia and influenza	78.6
Pneumonia	76.7
Influenza	1.9
Diabetes melitus	48.8
Suicide	30.8
Chronic liver disease and cirrhosis	25.6
Other infective and parasitic diseases	32.2
Human immunodeficiency virus (HIV) infections (AIDS)	24.1
Homicide and legal intervention	25.7
Nephritis, nephrotic syndrome, and nephrosis	20.9
Septicemia	19.8
Certain conditions originating in the perinatal period	17.5
Congenital anomalies	13.4
Benign neoplasms[3]	7.0
Ulcer of stomach and duodenum	6.2
Hernia of abdominal cavity and intestinal obstruction[4]	5.6
Anemias	4.2
Cholelithiasis and other disorders of gall bladder	3.0
Nutritional deficiencies	3.1
Tuberculosis	1.8
Infectionss of kidney	1.1
Viral hepatitis	1.7
Menengitis	1.2
Acute bronchitis and bronciolitis	0.6
Hyperplasia of prostate	0.3
Symptoms, signs, and ill-defined conditions	26.3
All other causes	174.1

Deaths classified according to ningth revision of *International Classification of Diseases*. Original source: U.S. National Center fo Health Statistics, *Vital Statistics of the United States*, annual; Monthly Vital Statistics Report and unpublished data.

[3] Includes neoplasms of unspecified nature and carcinoma in situ.
[4] Without mention of hernia