

DATA HORROR STORY: INFECTED DATA IN BIBLICAL TERMS



GENESIS

- First official case of Covid-19 in the Netherlands was on February the 27th
- RIVM is responsible for tracking the epidemic and advise government on actions in the Netherlands
- They started reporting data on their site at the beginning of March
- But they did it in an ancient way. They only published the summaries and not the raw data
- People wanting to use the data (me for instance) had a biblical path to get it and make it (re)usable so we could give others more insights

GENESIS

Wekelijkse update: 7 t/m 13 oktober 2020

	Afgelopen week ¹	Voorgaande week ²
Meldingen van COVID-19 door GGD'en		
Aantal nieuwe meldingen	43.903	27.485
Aantal meldingen ziekenhuisopnames op verpleegafdeling (bron: NICE)	1144	802
Aantal meldingen ziekenhuisopnames op Intensive Care (bron: NICE)	192	121
Overleden	150	89

GGD testlocaties per kalenderweek³	week 41	week 40
Totaal aantal afgenomen testen waarvan uitslag bekend is	239.639	223.274
Aantal positieve testen	33.038	23.264
Percentage positieve testen	13,8%	10,4%

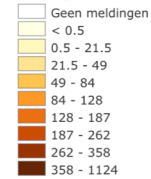
Niet alle patiënten zijn in de afgelopen week opgenomen in het ziekenhuis of overleden gemeld.

Covid-19 meldingen ▾

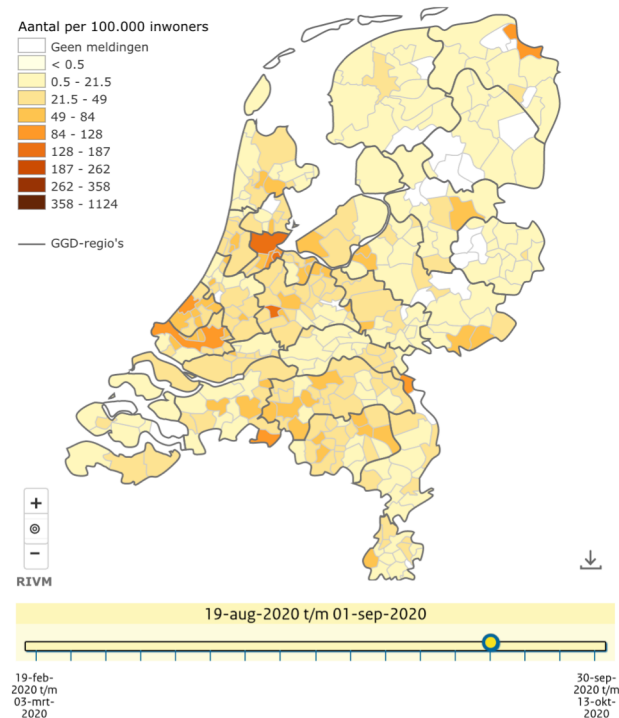
COVID-19 patiënten

Per gemeente van 19-aug-2020 t/m 01-sep-2020

Aantal per 100.000 inwoners



— GGD-regio's



Deze kaart toont over de afgelopen 2 weken, via een drop down menu, het aantal COVID-19 patiënten, het aantal vanwege COVID-19 in het ziekenhuis opgenomen patiënten, en het aantal overleden COVID-19 patiënten per 100.000 inwoners, zoals gemeld door de GGD'en. Deze gegevens worden elke dinsdag bijgewerkt.

Feedback

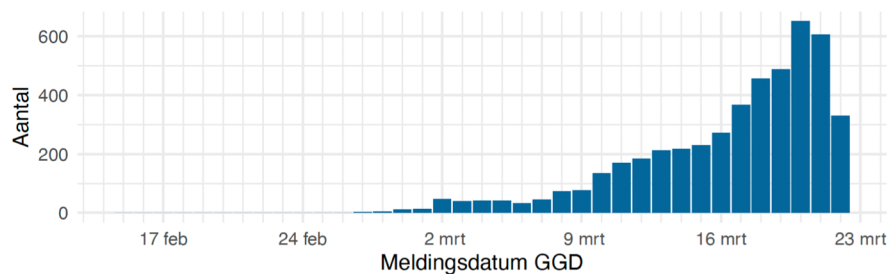
GENESIS

- Daily updates in HTML only for the first month
- Summaries, not full data
- No historic data, only latest
- First requests to RIVM to publish raw data as open data
- On March 23 the first "Epidemiologisch rapport" appeared with more details... in PDF!
- Needed to transcribe all relevant data every day of the week for several weeks

GENESIS

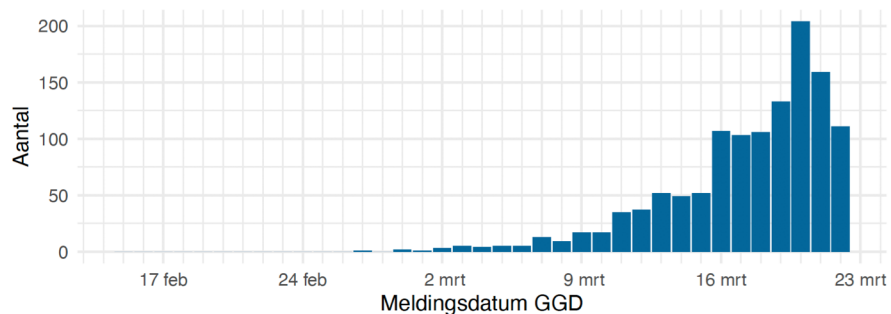
Aantal bij de GGD'en gemelde COVID-19 patiënten, naar meldingsdatum

Meldingen tot en met 22-03-2020.



Aantal bij de GGD'en gemelde COVID-19 patiënten opgenomen in het ziekenhuis, naar meldingsdatum

Meldingen tot en met 22-03-2020.



Tabel 4. Leeftijdsverdeling van bij de GGD'en gemelde COVID-19 patiënten, van patiënten in het ziekenhuis, en overleden patiënten².

Leeftijdsgroep	Totaal	%	Ziekenhuisopname	%	Overleden	%
Totaal gemeld	4749		1230		213	
0-4	14 (0.3)		8 (0.7)		0 (0)	
5-9	6 (0.1)		1 (0.1)		0 (0)	
10-14	27 (0.6)		2 (0.2)		0 (0)	
15-19	37 (0.8)		2 (0.2)		0 (0)	
20-24	90 (1.9)		3 (0.2)		0 (0)	
25-29	250 (5.3)		16 (1.3)		0 (0)	
30-34	267 (5.6)		16 (1.3)		0 (0)	
35-39	227 (4.8)		14 (1.1)		0 (0)	
40-44	209 (4.4)		13 (1.1)		0 (0)	
45-49	370 (7.8)		60 (4.9)		0 (0)	
50-54	383 (8.1)		73 (5.9)		0 (0)	
55-59	449 (9.5)		90 (7.3)		1 (0.5)	
60-64	367 (7.7)		106 (8.6)		5 (2.3)	
65-69	353 (7.4)		152 (12.4)		15 (7)	
70-74	403 (8.5)		168 (13.7)		20 (9.4)	
75-79	413 (8.7)		191 (15.5)		37 (17.4)	
80-84	409 (8.6)		161 (13.1)		66 (31)	
85-89	288 (6.1)		110 (8.9)		49 (23)	
90-94	127 (2.7)		36 (2.9)		14 (6.6)	
95+	37 (0.8)		5 (0.4)		5 (2.3)	
Niet bekend	23 (0.5)		3 (0.2)		1 (0.5)	

Tabel 5. Man-vrouwverdeling van bij de GGD'en gemelde COVID-19 patiënten, van in het ziekenhuis opgenomen COVID-19 patiënten, en van overleden COVID-19 patiënten.

GENESIS

- So somebody at RIVM had a lot of nice, usable data and put it in a table and used LaTeX to create a PDF
- Why not also publish the raw data?
- More requests and pleas to RIVM
- Meanwhile we copied the data manually and build our first PDF scrapper
- But...

Changing the layout and reference periods of the tables

18 April

Leeftijdverdeling en man-vrouwverdeling

Tabel 3. Leeftijdverdeling van bij de GGD'en gemelde COVID-19 patiënten, van in het ziekenhuis opgenomen COVID-19 patiënten en van overleden COVID-19 patiënten^{5,6}.

Leeftijdsgroep	Totaal	%	Ziekenhuisopname	%	Overleden	%
Totaal gemeld	31589		9504		3601	
0-4	65 (0.2)		37 (0.4)		0 (0.0)	
5-9	10 (0.0)		2 (0.0)		0 (0.0)	
10-14	42 (0.1)		6 (0.1)		0 (0.0)	
15-19	232 (0.7)		21 (0.2)		1 (0.0)	
20-24	1091 (3.5)		45 (0.5)		0 (0.0)	
25-29	1462 (4.6)		80 (0.9)		3 (0.1)	
30-34	1441 (4.6)		111 (1.2)		2 (0.1)	
35-39	1245 (3.9)		153 (1.6)		4 (0.1)	
40-44	1409 (4.5)		225 (2.3)		4 (0.1)	
45-49	2066 (6.5)		455 (4.7)		9 (0.2)	
50-54	2702 (8.6)		678 (7.1)		30 (0.8)	
55-59	3015 (9.5)		901 (9.4)		53 (1.5)	
60-64	2508 (7.9)		1045 (10.9)		104 (2.9)	
65-69	1880 (6.0)		1118 (11.7)		221 (6.1)	
70-74	2374 (7.5)		1444 (15.1)		395 (11.0)	
75-79	2637 (8.3)		1384 (14.4)		667 (18.5)	
80-84	2746 (8.7)		1020 (10.6)		796 (22.1)	
85-89	2653 (8.4)		641 (6.7)		768 (21.3)	
90-94	1517 (4.8)		185 (1.9)		405 (11.2)	
95+	479 (1.5)		33 (0.3)		138 (3.8)	
Niet vermeld	15 (0.0)		1 (0.0)		1 (0.0)	

6 May: change in layout

Leeftijdverdeling en man-vrouwverdeling

Tabel 3. Leeftijdverdeling van bij de GGD'en gemelde COVID-19 patiënten, van in het ziekenhuis opgenomen COVID-19 patiënten en van overleden COVID-19 patiënten^{5,6}.

Leeftijdsgroep	Totaal gemeld	%	Ziekenhuisopname	%	Overleden	%
Totaal gemeld	41319		11153		5204	
0-4	71 0.2		44 0.4		0 0.0	
5-9	15 0.0		2 0.0		0 0.0	
10-14	51 0.1		7 0.1		0 0.0	
15-19	406 1.0		27 0.2		1 0.0	
20-24	1656 4.0		52 0.5		0 0.0	
25-29	2030 4.9		102 0.9		3 0.1	
30-34	1908 4.6		145 1.3		3 0.1	
35-39	1672 4.0		184 1.6		6 0.1	
40-44	1906 4.6		267 2.4		5 0.1	
45-49	2795 6.8		539 4.8		17 0.3	
50-54	3605 8.7		803 7.2		37 0.7	
55-59	3988 9.7		1072 9.6		83 1.6	
60-64	3267 7.9		1223 11.0		144 2.8	
65-69	2182 5.3		1291 11.6		295 5.7	
70-74	2731 6.6		1627 14.6		550 10.6	
75-79	3125 7.6		1560 14.0		902 17.3	
80-84	3516 8.5		1174 10.5		1110 21.3	
85-89	3550 8.6		767 6.9		1144 22.0	
90-94	2133 5.2		228 2.0		659 12.7	
95+	704 1.7		38 0.3		244 4.7	
Niet vermeld	8 0.0		1 0.0		1 0.0	

June: change in reporting period

4 Leeftijdverdeling en man-vrouwverdeling van COVID-19 patiënten vanaf 4 mei 2020

Tabel 3: Leeftijdverdeling van bij de GGD'en gemelde COVID-19 patiënten, van in het ziekenhuis opgenomen COVID-19 patiënten en van overleden COVID-19 patiënten vanaf 4 mei 2020^{1,2}

Leeftijdsgroep	Totaal gemeld	%	Ziekenhuisopname	%	Overleden	%
Totaal gemeld	9015		436		783	
0-4	72 0.8		7 1.6		0 0.0	
5-9	82 0.9		0 0.0		0 0.0	
10-14	151 1.7		2 0.5		0 0.0	
15-19	289 3.2		3 0.7		0 0.0	
20-24	666 7.4		10 2.3		0 0.0	
25-29	765 8.5		9 2.1		0 0.0	
30-34	675 7.5		10 2.3		1 0.1	
35-39	553 6.1		9 2.1		1 0.1	
40-44	579 6.4		13 3.0		1 0.1	
45-49	686 7.6		32 7.3		6 0.8	
50-54	793 8.8		43 9.9		9 1.1	
55-59	789 8.8		58 13.3		14 1.8	
60-64	601 6.7		44 10.1		18 2.3	
65-69	288 3.2		40 9.2		34 4.3	
70-74	281 3.1		33 7.6		68 8.7	
75-79	320 3.5		37 8.5		87 11.1	
80-84	428 4.7		41 9.4		142 18.1	
85-89	534 5.9		30 6.9		201 25.7	
90-94	351 3.9		14 3.2		149 19.0	
95+	111 1.2		1 0.2		52 6.6	
Niet vermeld	1 0.0		0 0.0		0 0.0	

Switching order and adding columns

March

Provincie	Aantal	%
Totaal gemeld	4749	
Drenthe	49	1
Flevoland	67	1.4
Friesland	39	0.8
Gelderland	520	10.9
Groningen	71	1.5
Limburg	502	10.6
Noord-Brabant	1558	32.8
Noord-Holland	600	12.6
Overijssel	226	4.8
Utrecht	415	8.7
Zeeland	55	1.2
Zuid-Holland	647	13.6

April

Provincie	Aantal	%	Vershil met gisteren
Totaal gemeld	34134		729
Groningen	314	0.9	5
Friesland	475	1.4	6
Drenthe	403	1.2	4
Overijssel	2401	7.0	65
Flevoland	595	1.7	23
Gelderland	4356	12.8	98
Utrecht	2393	7.0	54
Noord-Holland	5033	14.7	99
Zuid-Holland	6886	20.2	144
Zeeland	532	1.6	4
Noord-Brabant	7209	21.1	140
Limburg	3537	10.4	87

GENESIS

- Position in PDF changed
- Table changed (dropped "()")
- Time period of table changed
- Order of rows changed
- Loads of manual checks and adjustments were needed to get the data right. Leading to errors

RESURRECTION



- On May 21st RIVM started publishing open data!
- So now we could use their daily updates as feed

	A	B	C	D
1				
2				
3	Rijlabels	Som van Total_reported	Som van Deceased	Som van Hospital_admission
4	13-mrt	804	9	115
5	14-mrt	959	12	136
6	15-mrt	1135	19	162
7	16-mrt	1413	23	205
8	17-mrt	1705	40	314
9	18-mrt	2051	56	408
10	19-mrt	2460	77	489
11	20-mrt	3651	105	648
12	21-mrt	3631	136	836
13	22-mrt	4204	180	988
14	23-mrt	4749	213	1230
15	24-mrt	5560	276	1495
16	25-mrt	6412	356	1836
17	26-mrt	7431	434	2151

	A	B	C	D	E	F	G
1	Date_of_report	Municipality_code	Municipality_name	Province	Total_reported	Hospital_admission	Deceased
2	13-03-2020 10:00	GM0003	Appingedam	Groningen	0	0	0
3	13-03-2020 10:00	GM0010	Delfzijl	Groningen	0	0	0
4	13-03-2020 10:00	GM0014	Groningen	Groningen	3	0	0
5	13-03-2020 10:00	GM0024	Loppersum	Groningen	0	0	0
6	13-03-2020 10:00	GM0034	Almere	Flevoland	1	1	0
7	13-03-2020 10:00	GM0037	Stadskanaal	Groningen	0	0	0
8	13-03-2020 10:00	GM0047	Veendam	Groningen	0	0	0
9	13-03-2020 10:00	GM0050	Zeevolde	Flevoland	1	0	0
10	13-03-2020 10:00	GM0059	Achtkarspelen	Friesland	0	0	0
11	13-03-2020 10:00	GM0060	Ameland	Friesland	0	0	0
12	13-03-2020 10:00	GM0072	Harlingen	Friesland	0	0	0
13	13-03-2020 10:00	GM0074	Heerenveen	Friesland	0	0	0
14	13-03-2020 10:00	GM0080	Leeuwarden	Friesland	1	0	0
15	13-03-2020 10:00	GM0085	Ooststellingwerf	Friesland	0	0	0
16	13-03-2020 10:00	GM0086	Opsterland	Friesland	2	0	0
17	13-03-2020 10:00	GM0088	Schiermonnikoog	Friesland	0	0	0
18	13-03-2020 10:00	GM0090	Smallingerland	Friesland	1	1	0
19	13-03-2020 10:00	GM0093	Terschelling	Friesland	0	0	0
20	13-03-2020 10:00	GM0096	Vlieland	Friesland	0	0	0

Rijksinstituut voor Volksgezondheid en Milieu
Ministerie van Volksgezondheid, Welzijn en Sport

COVID-19 dataset

Name	Last modified	Size	Description
 COVID-19_rioolwaterdata.json	18-Oct-2020 14:15	1.2M	JSON bestand
 COVID-19_rioolwaterdata.csv	18-Oct-2020 14:15	255K	Kommagescheiden bestand
 COVID-19_reproductiegetal.json	18-Oct-2020 14:15	19K	JSON bestand
 COVID-19_prevalentie.json	18-Oct-2020 14:15	21K	JSON bestand
 COVID-19_casus_landelijk.json	18-Oct-2020 14:15	58M	JSON bestand
 COVID-19_casus_landelijk.csv	18-Oct-2020 14:15	19M	Kommagescheiden bestand
 COVID-19_aantallen_gemeente_per_dag.json	19-Oct-2020 14:15	24M	JSON bestand
 COVID-19_aantallen_gemeente_per_dag.csv	19-Oct-2020 14:15	11M	Kommagescheiden bestand
 COVID-19_aantallen_gemeente_cumulatief.json	19-Oct-2020 14:15	14M	JSON bestand
 COVID-19_aantallen_gemeente_cumulatief.csv	19-Oct-2020 14:15	4.5M	Kommagescheiden bestand

RESURRECTION

- Now the data became way easier to use
- RIVM switched to weekly reports (still PDF), but others could use the daily updates and inform a wider audience in the days in between
- But strange and wonderful things happened
- In a period with a lot of people dying, it seemed the doctors had found a miracle cure: they could resurrect the death!

RESURRECTION

- On July 10th we had our first resurrection! And we even improved on the Bible by having two times two resurrections a couple of days later!

	A	B	C	D	E	F	G
1	Weeknummer	(Alle)					
2							
3	Rijlabels	Som van Deceased	Change				
120	07-jul	6132	4,00				
121	08-jul	6135	3,00				
122	09-jul	6137	2,00				
123	10-jul	6136	1,00				
124	11-jul	6137	1,00				
125	12-jul	6137	0,00				
126	13-jul	6137	0,00				
127	14-jul	6135	2,00				
128	15-jul	6136	1,00				
129	16-jul	6137	1,00				
130	17-jul	6138	1,00				
131	18-jul	6136	2,00				
132	19-jul	6136	0,00				
133	20-jul	6136	0,00				
134	21-jul	6136	0,00				
135	22-jul	6139	3,00				
136	23-jul	6139	0,00				
137	24-jul	6139	0,00				
138	25-jul	6140	1,00				

RESURRECTION

- Of course no real miracle happened...
- Those were administrative corrections on past data. Some deaths were found not to be Covid related and were removed from the lists
- But they didn't tell which one, or why
- We had to analyze the mismatch between old downloads and new to find what needed correction
- The other way happened as well. One day 20 deaths were added, but they happened weeks before. Made wrong headlines in the news for couple of hours before it was explained

BABYLONIAN DATA CONFUSION

- Using different date formats (often with meaningless time stamps)
- Introducing abbreviations without a readme for context (DOO, DPL, DON)
- Using different age groups

Leeftijdsverdeling en man-v

Tabel 3. Leeftijdsverdeling van bij de ziekenhuis opgenomen COVID-19 pat

Leeftijdsgroep	Totaal	%
Totaal gemeld	34134	
0-4	69	(0.2)
5-9	11	(0.0)
10-14	44	(0.1)
15-19	276	(0.8)
20-24	1255	(3.7)
25-29	1613	(4.7)
30-34	1555	(4.6)
35-39	1374	(4.0)
40-44	1520	(4.5)

	A	B	C	D
1	Date_file	Date_statistics	Date_statisti	Agegroup
2	17-10-2020 10:00	01-01-2020	DOO	40-49
3	17-10-2020 10:00	20-01-2020	DOO	50-59
4	17-10-2020 10:00	29-01-2020	DOO	80-89
5	17-10-2020 10:00	31-01-2020	DOO	80-89
6	17-10-2020 10:00	31-01-2020	DOO	90+
7	17-10-2020 10:00	01-02-2020	DOO	60-69
8	17-10-2020 10:00	01-02-2020	DOO	60-69
9	17-10-2020 10:00	01-02-2020	DOO	50-59
10	17-10-2020 10:00	03-02-2020	DOO	50-59
11	17-10-2020 10:00	03-02-2020	DOO	80-89
12	17-10-2020 10:00	06-02-2020	DOO	20-29
13	17-10-2020 10:00	06-02-2020	DOO	80-89
14	17-10-2020 10:00	07-02-2020	DOO	70-79

“INFECTED” DATA

DATA CONVERSATIONS

Let's talk open science
and research data!



- Hurts the trustworthiness of the RIVM
- Created confusion in the public domain in already challenging times
- Frustrated researchers, journalists and data analysts. And consumes valuable time

So be better data prepared of
the next (biblical) plague!