

Table 1. Codes for defining population characteristics and healthcare utilisation

Data source and variable	Code		
Danish Civil Registration System			
Municipality	Municipality code		
Hedensted	766		
Horsens	615		
Odder	727		
Skanderborg	746		
Danish Register for Evaluation of Marginalization			
Unemployed	111-113, 211-218, 299		
Sick leave	890, 893-899		
Educational grants	140, 143-149, 151, 231-232, 651-652, 661, 720, 723-729		
Maternity or paternity leave	412, 881		
Retirement pension	621, 998		
Disability pension	783		
Cash benefits	130, 133-139, 710, 713-718, 730, 733-739		
Danish National Health Service Register			
	Specialty	Service	Time
GP daytime face-to-face contacts	80	0101	0,1
GP out-of-hours face-to-face contacts	8X	0101	8,9
GP blood sample tests	80	2601, 2101	Any
GP hemoglobin measuring tests	80	7108	Any
GP spirometer tests	80	7113, 7121, 7183	Any
GP electrocardiographs	80	7156	Any
GP talk therapy sessions	80	6101, 4003, 4021, 4022, 4023, 4024, 4025, 4026, 4027, 4050, 4063, 4106, 4247, 4248, 4249	Any
GP urine sample tests	80	7101	Any
GP rapid strep tests	80	7109	Any
GP C-reactive protein tests	80	7120	Any
Other GP services	80	Any other	Any
Chiropractor services	53, 64	Any	Any
Dentist services	50	Any	Any
Physiotherapist services	51, 61, 62	Any	Any
Podiatrist services	54	Any	Any
Psychologist services	24, 35, 63	Any	Any
Danish National Database of Reimbursed Prescriptions			
	ATC codes		
Antibiotics	J01		
Antidiabetics	A10A, A10B		
Aspirins	B01AC06, N02BA01, N02BA51		
Blood pressure lowering medication	C02, C07, C08, C09		
Statins	C10AA		
Proton pump inhibitors	A02BC		
Systemic glucocorticoids	H02AB		
Inhaled corticosteroid therapy	R03		
NSAIDs	M01A		
Opioids	N02A		
Antidepressants	N06A		
Antipsychotics	N05A		

Prehospital electronic record	Department code	
Contact with the prehospital	6610	
Emergency call	6610020	
Medical car	6610040	
Helicopter	6610010	
Danish National Patient Registry		
Charlson comorbidity score	ICD 10 codes	Score
Myocardial infarction	I21, I22, I23	1
Congestive heart failure	I50, I110, I130, I132	1
Peripheral vascular disease	I70, I71, I72, I73, I74, I77	1
Cerebrovascular disease	I60-I69, G45, G46	1
Dementia	F00-F03, F05.1, G30	1
Chronic pulmonary disease	J40-J47, J60-J67, J68.4, J70.1, J70.3, J84.1, J92.0, J96.1, J98.2, J98.3	1
Connective tissue disease	M05, M06, M08, M09, M30, M31, M32, M33, M34, M35, M36, D86	1
Ulcer disease	K22.1, K25-K28	1
Mild liver disease	B18, K70.0-K70.3, K70.9, K71, K73, K74, K76.0	1
Diabetes	E10.0, E10.1, E10.9, E11.0, E11.1, E11.9	1
Hemiplegia	G81, G82	2
Moderate to severe renal disease	I12, I13, N00-N05, N07, N11, N14, N17-N19, Q61	2
Diabetes with end organ damage	E10.2-E10.8, E11.2-E11.8	2
Any tumor	C00-C75	2
Leukemia	C91-C95	2
Lymphoma	C81-C85, C88, C90, C96	2
Moderate to severe liver disease	B15.0, B16.0, B16.2, B19.0, K70.4, K72, K76.6, I85	3
Metastatic solid tumor	C76-C80	6
AIDS	B21-B24	6
Potentially preventable admissions	Acute contact with a main diagnosis among the following ICD 10 codes: I61, I63, I64, E869, K590, J12-J18, N30 except N304, J40-J47, I20-I25, A09, S22, S32, S42, S62, S72, S82, S92, D50-D53, M05-M07, M10-M13, M15-M19	
Electronic health record	Procedure codes	
Measuring patient's weight (kg)	ZZ0240	
Measuring patient's height (cm)	ZZ0241	
Determination of patient's body mass index	ZZ0242	

Table 2. Cleaning of body mass index registrations from the electronic health record

The electronic health record (EHR) contains registrations of patient’s height and weight used for calculating body mass index (BMI). Each registration is accompanied by a date and time stamp of the registration. When a patient’s height and weight are entered, the corresponding BMI can be calculated when the health professional clicks a button. Registrations are modifiable but everything is logged.

In this data cleaning algorithm, we focused on registrations of height and weight. We applied a series of exclusions and cleaning of registrations of height and weight. In case of multiple registrations in one patient on the same date and time, we included the most likely value (see details below).

An erroneous registration of height or weight resulted in an erroneous BMI calculation, hence BMI calculations corresponding to excluded or cleaned registrations were also excluded. The BMI calculations that were not excluded were compared to BMI calculated based on the cleaned registrations of height and weight.

Description of data cleaning	Number cleaned	Number excluded	Number included			
			Height	Weight	BMI	Total
Total registrations in the EHR	-	-	384,553	618,924	334,614	1,338,091
Patients aged <18 years	-	26,938	375,488	607,638	328,027	1,311,153
BMI<=0	-	28	375,488	607,638	327,999	1,311,125
Multiple registrations of same value on same date and time	-	12,229	374,033	603,975	320,888	1,298,896
Height <=1 cm	-	304	373,729	603,975	320,888	1,298,592
Weight <=0 kg	-	1,348	373,729	602,627	320,888	1,297,244
Height >1 to <2 cm was assumed to be in meters	149	-	373,729	602,627	320,886	1,297,242
Height < 120 cm combined with a higher and more likely height on the same date and time	-	6	373,723	602,627	320,888	1,297,238
Weight > 150 kg combined with a lower and more likely weight on the same date and time	-	8	373,723	602,619	320,888	1,297,230
Height < 120 cm combined with a weight > 150 kg are assumed to be switched	496	-	373,723	602,619	320,888	1,297,230
Height < 120 cm	-	451	373,272	602,619	320,888	1,296,779
Height > 250 cm	-	15	373,257	602,619	320,888	1,296,764
Weight < 30 kg for males	-	47	373,257	602,572	320,888	1,296,717
Weight < 20 kg for females	-	39	373,257	602,533	320,888	1,296,678
Weight > 235 kg	-	43	373,257	602,490	320,888	1,296,635
Corresponding erroneous BMI calculations	-	448	373,257	602,490	320,440	1,296,187

Estimating patient’s height

As we only included patients aged 18+ years, we assumed that their height did not change much during the inclusion period. After the initial data cleaning we estimated each patient’s height based on registrations:

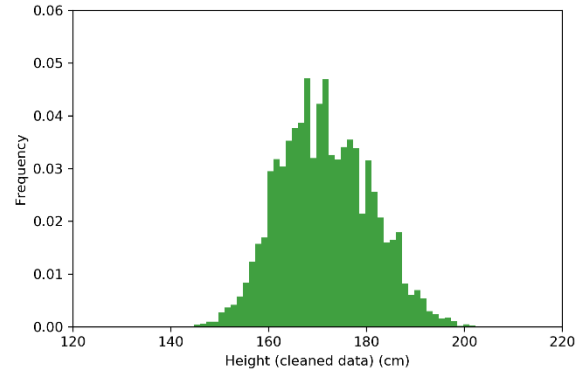
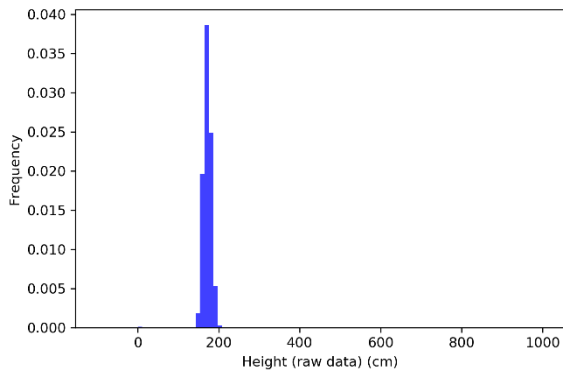
- For each patient, we calculated minimum, maximum, *preliminary mean*, and the number of registrations of height.
- For patients with at least five registrations of height, we identified outliers as registrations that differ $\pm >8$ cm from their *preliminary mean* among patients with a difference of >10 cm between their minimum and maximum height registration. We excluded 2,081 outliers and 1,823 corresponding erroneous BMI calculations among 1,433 patients.
- After exclusion of outliers, each patient’s mean height was calculated.
- Patient’s mean height was used to calculate BMI at each registration of patient’s weight.

The final cleaned data set consisted of 601,999 BMI calculations among 101,507 patients.

Comparisons of raw and cleaned data

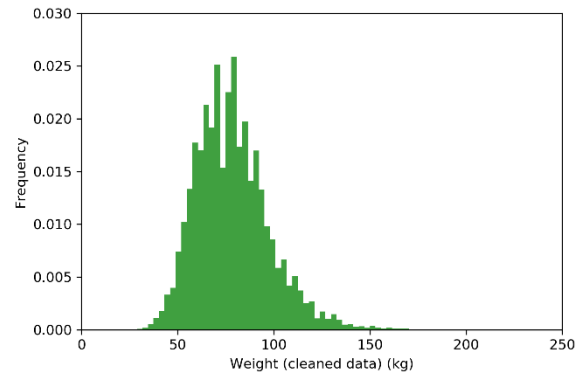
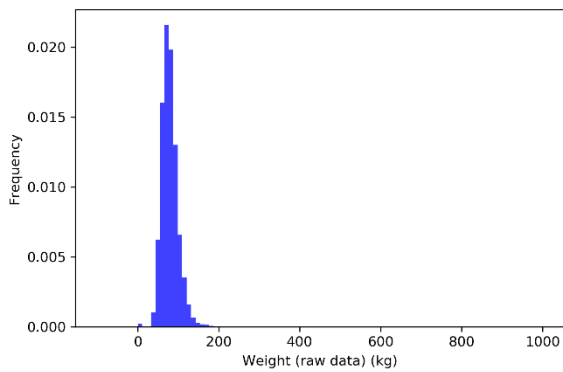
Height (cm)

	Mean	SD	Min	5%	25%	50%	75%	95%	Max
Raw data	171	13	-99	157	165	171	178	187	999
Cleaned data	171	9	129	157	165	171	178	187	216



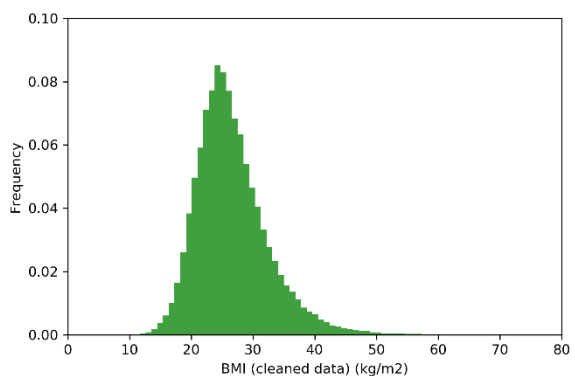
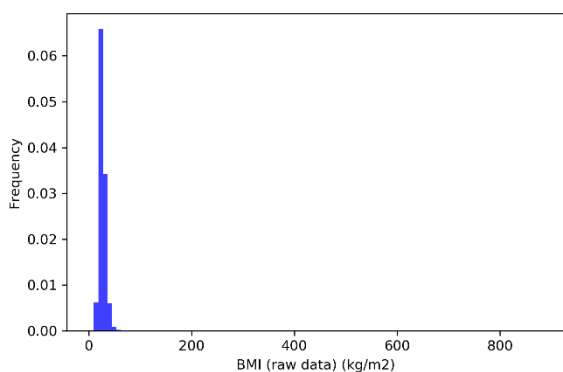
Weight (kg)

	Mean	SD	Min	5%	25%	50%	75%	95%	Max
Raw data	79	21	-99	51	65	77	90	115	999
Cleaned data	79	20	23	51	65	76	90	114	225



BMI (kg/m²)

	Mean	SD	Min	5%	25%	50%	75%	95%	Max
Raw data	26.9	11.9	1.2	18.7	22.7	25.7	29.7	37.6	881
Cleaned data	26.6	5.9	9.8	18.7	22.7	25.7	29.6	37.4	103



Bland-Altman plot comparing raw and cleaned BMI

	Mean	SD	95% Limits of agreement
Difference raw - cleaned data (kg/m ²)	0.0034	0.6129	(-1.1979, 1.2047)

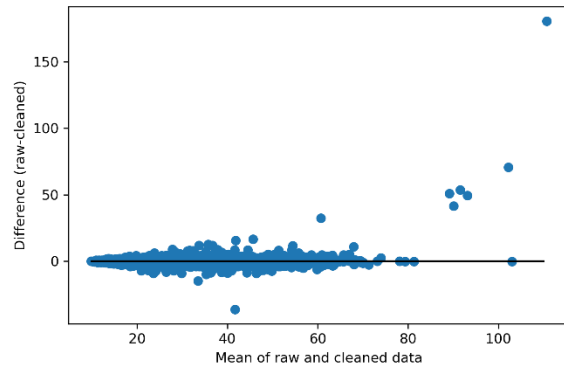
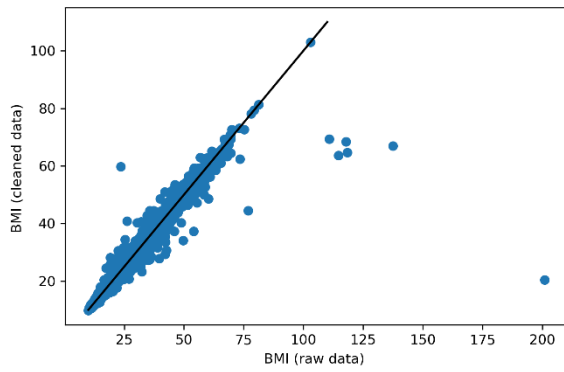


Table 3. Cleaning of blood pressure registrations from the electronic health record

The electronic health record (EHR) contains registrations of patient’s diastolic and systolic blood pressure. Registrations are accompanied by a date and time stamp. Registrations are modifiable but everything is logged.

Description of data cleaning	Number cleaned	Number excluded	Number included		
			Diastolic	Systolic	Total
Total registrations in the EHR	-	-	1,987,416	1,991,026	3,978,442
Multiple registrations of same value on same date and time	-	11,931	1,980,836	1,985,675	3,966,511
Diastolic blood pressure <30 or >250 mm Hg	-	2,270	1,978,566	1,985,675	3,964,241
Systolic blood pressure <30 or >250 mm Hg	-	2,152	1,978,566	1,983,523	3,962,089

Based on these 1,978,566 diastolic and 1,983,523 systolic blood pressure registrations, there were 1,935,103 combined registrations of diastolic and systolic blood pressure on the same date and time among 105,874 patients. 1,902,502 (98%) of the combined registrations consisted of one measurement of diastolic and systolic blood pressure, whereas 32,601 (2%) consisted of multiple registrations in one patient on the same date and time. 32,168 (99%) of these multiple registrations consisted of up to three registrations of diastolic and/or systolic blood pressure. Depending on the aim of the data extraction, you could choose to include only the lowest or highest value registered among these multiple registrations. We included all 2,087,892 combined registrations of diastolic and systolic blood pressure on the same date and time.

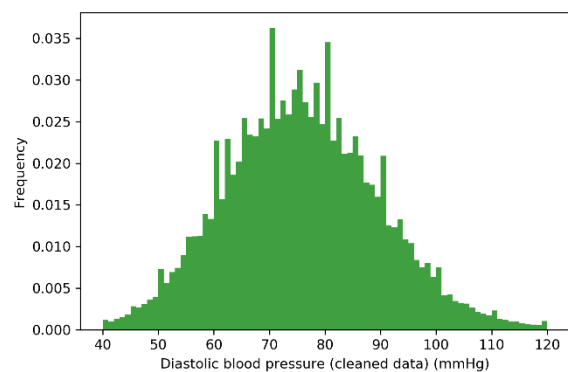
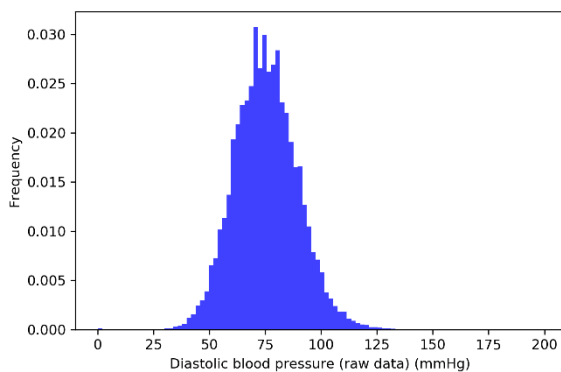
Description of data cleaning	Number cleaned	Number excluded	Number included
Total combinations of diastolic and systolic blood pressure in the EHR	-	-	2,087,892
Diastolic blood pressure > systolic blood pressure	831	-	2,087,892
Diastolic blood pressure <40 or >120 mm Hg	-	14,027	2,073,865
Systolic blood pressure <75 or >220 mm Hg	-	5,882	2,067,983

The final cleaned data set consisted of 2,067,983 combined registrations of diastolic and systolic blood pressure among 105,766 patients.

Comparisons of raw and cleaned data

Diastolic blood pressure (mm Hg)

	Mean	SD	Min	5%	25%	50%	75%	95%	Max
Raw data	77	101	-136	53	65	75	85	99	9988
Cleaned data	75	14	40	53	66	75	85	99	120



Systolic blood pressure (mm Hg)

	Mean	SD	Min	5%	25%	50%	75%	95%	Max
Raw data	132	60	0	97	115	129	145	174	9999
Cleaned data	132	23	75	98	115	130	145	173	220

