lab-06

AUTHOR
Andre Gala-Garza

PUBLISHED
February 19, 2026

Disclaimer: Generative AI was used to assist with templating and writing code in this assignment; however, this code was checked manually and edited by hand to ensure accuracy.

Source: OpenAI. (2026). *ChatGPT (GPT-5.2 Thinking)* [Large language model]. https://chatgpt.com/.

0. Load packages and data

```
library("dplyr")
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library("ggplot2")
library("tidytext")
library("tidyr")
library("forcats")
library("stringr")
```

```
url <- 'https://raw.githubusercontent.com/dmcable/BIOSTAT620W26/main/data/mtsamples/mtsamples.csv
if (!file.exists("mtsamples.csv"))
  download.file(
    url = url,
    destfile = "mtsamples.csv",
    method   = "libcurl",
    timeout  = 60
    )
mtsamples <- read.csv('mtsamples.csv')
```

```
dim(mtsamples)
```

```
[1] 4999    6
```

```r
head(mtsamples)
```

```
  X
1 0
2 1
3 2
4 3
5 4
6 5

description
1
A 23-year-old white female presents with complaint of allergies.
2
Consult for laparoscopic gastric bypass.
3
Consult for laparoscopic gastric bypass.
4
2-D M-Mode. Doppler.
5
2-D Echocardiogram
6  Morbid obesity.  Laparoscopic antecolic antegastric Roux-en-Y gastric bypass with EEA
anastomosis.  This is a 30-year-old female, who has been overweight for many years.  She has
tried many different diets, but is unsuccessful.
            medical_specialty                               sample_name
1        Allergy / Immunology                        Allergic Rhinitis
2                  Bariatrics  Laparoscopic Gastric Bypass Consult - 2
3                  Bariatrics  Laparoscopic Gastric Bypass Consult - 1
4  Cardiovascular / Pulmonary                   2-D Echocardiogram - 1
5  Cardiovascular / Pulmonary                   2-D Echocardiogram - 2
6                  Bariatrics              Laparoscopic Gastric Bypass

transcription
1
SUBJECTIVE:,  This 23-year-old white female presents with complaint of allergies.  She used to
have allergies when she lived in Seattle but she thinks they are worse here.  In the past, she
has tried Claritin, and Zyrtec.  Both worked for short time but then seemed to lose
effectiveness.  She has used Allegra also.  She used that last summer and she began using it
again two weeks ago.  It does not appear to be working very well.  She has used over-the-counter
sprays but no prescription nasal sprays.  She does have asthma but doest not require daily
medication for this and does not think it is flaring up.,MEDICATIONS: , Her only medication
currently is Ortho Tri-Cyclen and the Allegra.,ALLERGIES: , She has no known medicine
allergies.,OBJECTIVE:,Vitals:  Weight was 130 pounds and blood pressure 124/78.,HEENT:  Her
throat was mildly erythematous without exudate.  Nasal mucosa was erythematous and swollen.  Only
clear drainage was seen.  TMs were clear.,Neck:  Supple without adenopathy.,Lungs:
Clear.,ASSESSMENT:,  Allergic rhinitis.,PLAN:,1.  She will try Zyrtec instead of Allegra again.
Another option will be to use loratadine.  She does not think she has prescription coverage so
that might be cheaper.,2.  Samples of Nasonex two sprays in each nostril given for three weeks.
A prescription was written as well.
2
```

PAST MEDICAL HISTORY:, He has difficulty climbing stairs, difficulty with airline seats, tying shoes, used to public seating, and lifting objects off the floor. He exercises three times a week at home and does cardio. He has difficulty walking two blocks or five flights of stairs. Difficulty with snoring. He has muscle and joint pains including knee pain, back pain, foot and ankle pain, and swelling. He has gastroesophageal reflux disease.,PAST SURGICAL HISTORY:, Includes reconstructive surgery on his right hand 13 years ago. ,SOCIAL HISTORY:, He is currently single. He has about ten drinks a year. He had smoked significantly up until several months ago. He now smokes less than three cigarettes a day.,FAMILY HISTORY:, Heart disease in both grandfathers, grandmother with stroke, and a grandmother with diabetes. Denies obesity and hypertension in other family members.,CURRENT MEDICATIONS:, None.,ALLERGIES:, He is allergic to Penicillin.,MISCELLANEOUS/EATING HISTORY:, He has been going to support groups for seven months with Lynn Holmberg in Greenwich and he is from Eastchester, New York and he feels that we are the appropriate program. He had a poor experience with the Greenwich program. Eating history, he is not an emotional eater. Does not like sweets. He likes big portions and carbohydrates. He likes chicken and not steak. He currently weighs 312 pounds. Ideal body weight would be 170 pounds. He is 142 pounds overweight. If ,he lost 60% of his excess body weight that would be 84 pounds and he should weigh about 228.,REVIEW OF SYSTEMS: ,Negative for head, neck, heart, lungs, GI, GU, orthopedic, and skin. Specifically denies chest pain, heart attack, coronary artery disease, congestive heart failure, arrhythmia, atrial fibrillation, pacemaker, high cholesterol, pulmonary embolism, high blood pressure, CVA, venous insufficiency, thrombophlebitis, asthma, shortness of breath, COPD, emphysema, sleep apnea, diabetes, leg and foot swelling, osteoarthritis, rheumatoid arthritis, hiatal hernia, peptic ulcer disease, gallstones, infected gallbladder, pancreatitis, fatty liver, hepatitis, hemorrhoids, rectal bleeding, polyps, incontinence of stool, urinary stress incontinence, or cancer. Denies cellulitis, pseudotumor cerebri, meningitis, or encephalitis.,PHYSICAL EXAMINATION:, He is alert and oriented x 3. Cranial nerves II-XII are intact. Afebrile. Vital Signs are stable.

3 HISTORY OF PRESENT ILLNESS: , I have seen ABC today. He is a very pleasant gentleman who is 42 years old, 344 pounds. He is 5'9". He has a BMI of 51. He has been overweight for ten years since the age of 33, at his highest he was 358 pounds, at his lowest 260. He is pursuing surgical attempts of weight loss to feel good, get healthy, and begin to exercise again. He wants to be able to exercise and play volleyball. Physically, he is sluggish. He gets tired quickly. He does not go out often. When he loses weight he always regains it and he gains back more than he lost. His biggest weight loss is 25 pounds and it was three months before he gained it back. He did six months of not drinking alcohol and not taking in many calories. He has been on multiple commercial weight loss programs including Slim Fast for one month one year ago and Atkin's Diet for one month two years ago.,PAST MEDICAL HISTORY: , He has difficulty climbing stairs, difficulty with airline seats, tying shoes, used to public seating, difficulty walking, high cholesterol, and high blood pressure. He has asthma and difficulty walking two blocks or going eight to ten steps. He has sleep apnea and snoring. He is a diabetic, on medication. He has joint pain, knee pain, back pain, foot and ankle pain, leg and foot swelling. He has hemorrhoids.,PAST SURGICAL HISTORY: , Includes orthopedic or knee surgery.,SOCIAL HISTORY: , He is currently single. He drinks alcohol ten to twelve drinks a week, but does not drink five days a week and then will binge drink. He smokes one and a half pack a day for 15 years, but he has recently stopped smoking for the past two weeks.,FAMILY HISTORY: , Obesity, heart disease, and diabetes. Family history is negative for hypertension and stroke.,CURRENT MEDICATIONS:, Include Diovan, Crestor, and Tricor.,MISCELLANEOUS/EATING HISTORY: ,He says a couple of friends of his have had heart attacks and have had died. He used to drink everyday, but stopped two years ago. He now only drinks on weekends. He is on his second week of Chantix, which is a medication to come off smoking completely. Eating, he eats bad food. He is single. He eats things like bacon, eggs, and cheese, cheeseburgers, fast food, eats four times a day, seven in the morning, at noon, 9 p.m., and 2 a.m. He currently weighs 344 pounds and 5'9". His ideal body weight is

160 pounds. He is 184 pounds overweight. If he lost 70% of his excess body weight that would be 129 pounds and that would get him down to 215.,REVIEW OF SYSTEMS: , Negative for head, neck, heart, lungs, GI, GU, orthopedic, or skin. He also is positive for gout. He denies chest pain, heart attack, coronary artery disease, congestive heart failure, arrhythmia, atrial fibrillation, pacemaker, pulmonary embolism, or CVA. He denies venous insufficiency or thrombophlebitis. Denies shortness of breath, COPD, or emphysema. Denies thyroid problems, hip pain, osteoarthritis, rheumatoid arthritis, GERD, hiatal hernia, peptic ulcer disease, gallstones, infected gallbladder, pancreatitis, fatty liver, hepatitis, rectal bleeding, polyps, incontinence of stool, urinary stress incontinence, or cancer. He denies cellulitis, pseudotumor cerebri, meningitis, or encephalitis.,PHYSICAL EXAMINATION: ,He is alert and oriented x 3. Cranial nerves II-XII are intact. Neck is soft and supple. Lungs: He has positive wheezing bilaterally. Heart is regular rhythm and rate. His abdomen is soft. Extremities: He has 1+ pitting edema.,IMPRESSION/PLAN:, I have explained to him the risks and potential complications of laparoscopic gastric bypass in detail and these include bleeding, infection, deep venous thrombosis, pulmonary embolism, leakage from the gastrojejuno-anastomosis, jejunojejuno-anastomosis, and possible bowel obstruction among other potential complications. He understands. He wants to proceed with workup and evaluation for laparoscopic Roux-en-Y gastric bypass. He will need to get a letter of approval from Dr. XYZ. He will need to see a nutritionist and mental health worker. He will need an upper endoscopy by either Dr. XYZ. He will need to go to Dr. XYZ as he previously had a sleep study. We will need another sleep study. He will need H. pylori testing, thyroid function tests, LFTs, glycosylated hemoglobin, and fasting blood sugar. After this is performed, we will submit him for insurance approval.

4

2-D M-MODE: , ,1. Left atrial enlargement with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left ventricular ejection fraction of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology of aortic valve, mitral valve, tricuspid valve, and pulmonary valve.,7. PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral and tricuspid regurgitation.,2. Trace aortic and pulmonary regurgitation.

5

1. The left ventricular cavity size and wall thickness appear normal. The wall motion and left ventricular systolic function appears hyperdynamic with estimated ejection fraction of 70% to 75%. There is near-cavity obliteration seen. There also appears to be increased left ventricular outflow tract gradient at the mid cavity level consistent with hyperdynamic left ventricular systolic function. There is abnormal left ventricular relaxation pattern seen as well as elevated left atrial pressures seen by Doppler examination.,2. The left atrium appears mildly dilated.,3. The right atrium and right ventricle appear normal.,4. The aortic root appears normal.,5. The aortic valve appears calcified with mild aortic valve stenosis, calculated aortic valve area is 1.3 cm square with a maximum instantaneous gradient of 34 and a mean gradient of 19 mm.,6. There is mitral annular calcification extending to leaflets and supportive structures with thickening of mitral valve leaflets with mild mitral regurgitation.,7. The tricuspid valve appears normal with trace tricuspid regurgitation with moderate pulmonary artery hypertension. Estimated pulmonary artery systolic pressure is 49 mmHg. Estimated right atrial pressure of 10 mmHg.,8. The pulmonary valve appears normal with trace pulmonary insufficiency.,9. There is no pericardial effusion or intracardiac mass seen.,10. There is a color Doppler suggestive of a patent foramen ovale with lipomatous hypertrophy of the interatrial septum.,11. The study was somewhat technically limited and hence subtle abnormalities could be missed from the study.,

6

PREOPERATIVE DIAGNOSIS: , Morbid obesity.,POSTOPERATIVE DIAGNOSIS: ,Morbid obesity.,PROCEDURE: , Laparoscopic antecolic antegastric Roux-en-Y gastric bypass with EEA anastomosis.,ANESTHESIA: , General with endotracheal intubation.,INDICATION FOR PROCEDURE: , This is a 30-year-old female, who has been overweight for many years. She has tried many different diets, but is unsuccessful. She has been to our Bariatric Surgery Seminar, received some handouts, and signed the consent. The risks and benefits of the procedure have been explained to the patient.,PROCEDURE IN DETAIL: ,The patient was taken to the operating room and placed supine on the operating room table. All pressure points were carefully padded. She was given general anesthesia with endotracheal intubation. SCD stockings were placed on both legs. Foley catheter was placed for bladder decompression. The abdomen was then prepped and draped in standard sterile surgical fashion. Marcaine was then injected through umbilicus. A small incision was made. A Veress needle was introduced into the abdomen. CO2 insufflation was done to a maximum pressure of 15 mmHg. A 12-mm VersaStep port was placed through the umbilicus. I then placed a 5-mm port just anterior to the midaxillary line and just subcostal on the right side. I placed another 5-mm port in the midclavicular line just subcostal on the right side, a few centimeters below and medial to that, I placed a 12-mm VersaStep port. On the left side, just anterior to the midaxillary line and just subcostal, I placed a 5-mm port. A few centimeters below and medial to that, I placed a 15-mm port. I began by lifting up the omentum and identifying the transverse colon and lifting that up and thereby identifying my ligament of Treitz. I ran the small bowel down approximately 40 cm and divided the small bowel with a white load GIA stapler. I then divided the mesentery all the way down to the base of the mesentery with a LigaSure device. I then ran the distal bowel down, approximately 100 cm, and at 100 cm, I made a hole at the antimesenteric portion of the Roux limb and a hole in the antimesenteric portion of the duodenogastric limb, and I passed a 45 white load stapler and fired a stapler creating a side-to-side anastomosis. I reapproximated the edges of the defect. I lifted it up and stapled across it with another white load stapler. I then closed the mesenteric defect with interrupted Surgidac sutures. I divided the omentum all the way down to the colon in order to create a passageway for my small bowel to go antecolic. I then put the patient in reverse Trendelenburg. I placed a liver retractor, identified, and dissected the angle of His. I then dissected on the lesser curve, approximately 2.5 cm below the gastroesophageal junction, and got into a lesser space. I fired transversely across the stomach with a 45 blue load stapler. I then used two fires of the 60 blue load with SeamGuard to go up into my angle of His, thereby creating my gastric pouch. I then made a hole at the base of the gastric pouch and had Anesthesia remove the bougie and place the OG tube connected to the anvil. I pulled the anvil into place, and I then opened up my 15-mm port site and passed my EEA stapler. I passed that in the end of my Roux limb and had the spike come out antimesenteric. I joined the spike with the anvil and fired a stapler creating an end-to-side anastomosis, then divided across the redundant portion of my Roux limb with a white load GI stapler, and removed it with an Endocatch bag. I put some additional 2-0 Vicryl sutures in the anastomosis for further security. I then placed a bowel clamp across the bowel. I went above and passed an EGD scope into the mouth down to the esophagus and into the gastric pouch. I distended gastric pouch with air. There was no air leak seen. I could pass the scope easily through the anastomosis. There was no bleeding seen through the scope. We closed the 15-mm port site with interrupted 0 Vicryl suture utilizing Carter-Thomason. I copiously irrigated out that incision with about 2 L of saline. I then closed the skin of all incisions with running Monocryl. Sponge, instrument, and needle counts were correct at the end of the case. The patient tolerated the procedure well without any complications.

keywords

1

allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal, erythematous, allegra, sprays, allergic,

2

bariatrics, laparoscopic gastric bypass, weight loss programs, gastric bypass, atkin's diet,

```
weight watcher's, body weight, laparoscopic gastric, weight loss, pounds, months, weight,
laparoscopic, band, loss, diets, overweight, lost
3
bariatrics, laparoscopic gastric bypass, heart attacks, body weight, pulmonary embolism,
potential complications, sleep study, weight loss, gastric bypass, anastomosis, loss, sleep,
laparoscopic, gastric, bypass, heart, pounds, weight,
4                                                                           cardiovascular /
pulmonary, 2-d m-mode, doppler, aortic valve, atrial enlargement, diastolic function, ejection
fraction, mitral, mitral valve, pericardial effusion, pulmonary valve, regurgitation, systolic
function, tricuspid, tricuspid valve, normal lv
5 cardiovascular / pulmonary, 2-d, doppler, echocardiogram, annular, aortic root, aortic valve,
atrial, atrium, calcification, cavity, ejection fraction, mitral, obliteration, outflow,
regurgitation, relaxation pattern, stenosis, systolic function, tricuspid, valve, ventricular,
ventricular cavity, wall motion, pulmonary artery
6
bariatrics, gastric bypass, eea anastomosis, roux-en-y, antegastric, antecolic, morbid obesity,
roux limb, gastric pouch, intubation, laparoscopic, bypass, roux, endotracheal, anastomosis,
gastric
```

```r
colnames(mtsamples)
```

```
[1] "X"                 "description"       "medical_specialty"
[4] "sample_name"       "transcription"     "keywords"
```

```r
str(mtsamples)
```

```
'data.frame':   4999 obs. of  6 variables:
 $ X                : int  0 1 2 3 4 5 6 7 8 9 ...
 $ description      : chr  " A 23-year-old white female presents with complaint of allergies." "
Consult for laparoscopic gastric bypass." " Consult for laparoscopic gastric bypass." " 2-D M-
Mode. Doppler.  " ...
 $ medical_specialty: chr  " Allergy / Immunology" " Bariatrics" " Bariatrics" " Cardiovascular /
Pulmonary" ...
 $ sample_name      : chr  " Allergic Rhinitis " " Laparoscopic Gastric Bypass Consult - 2 " "
Laparoscopic Gastric Bypass Consult - 1 " " 2-D Echocardiogram - 1 " ...
 $ transcription    : chr  "SUBJECTIVE:,  This 23-year-old white female presents with complaint
of allergies.  She used to have allergies w"| __truncated__ "PAST MEDICAL HISTORY:, He has
difficulty climbing stairs, difficulty with airline seats, tying shoes, used to p"| __truncated__
"HISTORY OF PRESENT ILLNESS: , I have seen ABC today.  He is a very pleasant gentleman who is 42
years old, 344 "| __truncated__ "2-D M-MODE: , ,1.  Left atrial enlargement with left atrial
diameter of 4.7 cm.,2.  Normal size right and left "| __truncated__ ...
 $ keywords         : chr  "allergy / immunology, allergic rhinitis, allergies, asthma, nasal
sprays, rhinitis, nasal, erythematous, allegr"| __truncated__ "bariatrics, laparoscopic gastric
bypass, weight loss programs, gastric bypass, atkin's diet, weight watcher's, "| __truncated__
"bariatrics, laparoscopic gastric bypass, heart attacks, body weight, pulmonary embolism,
potential complication"| __truncated__ "cardiovascular / pulmonary, 2-d m-mode, doppler, aortic
valve, atrial enlargement, diastolic function, ejection"| __truncated__ ...
```

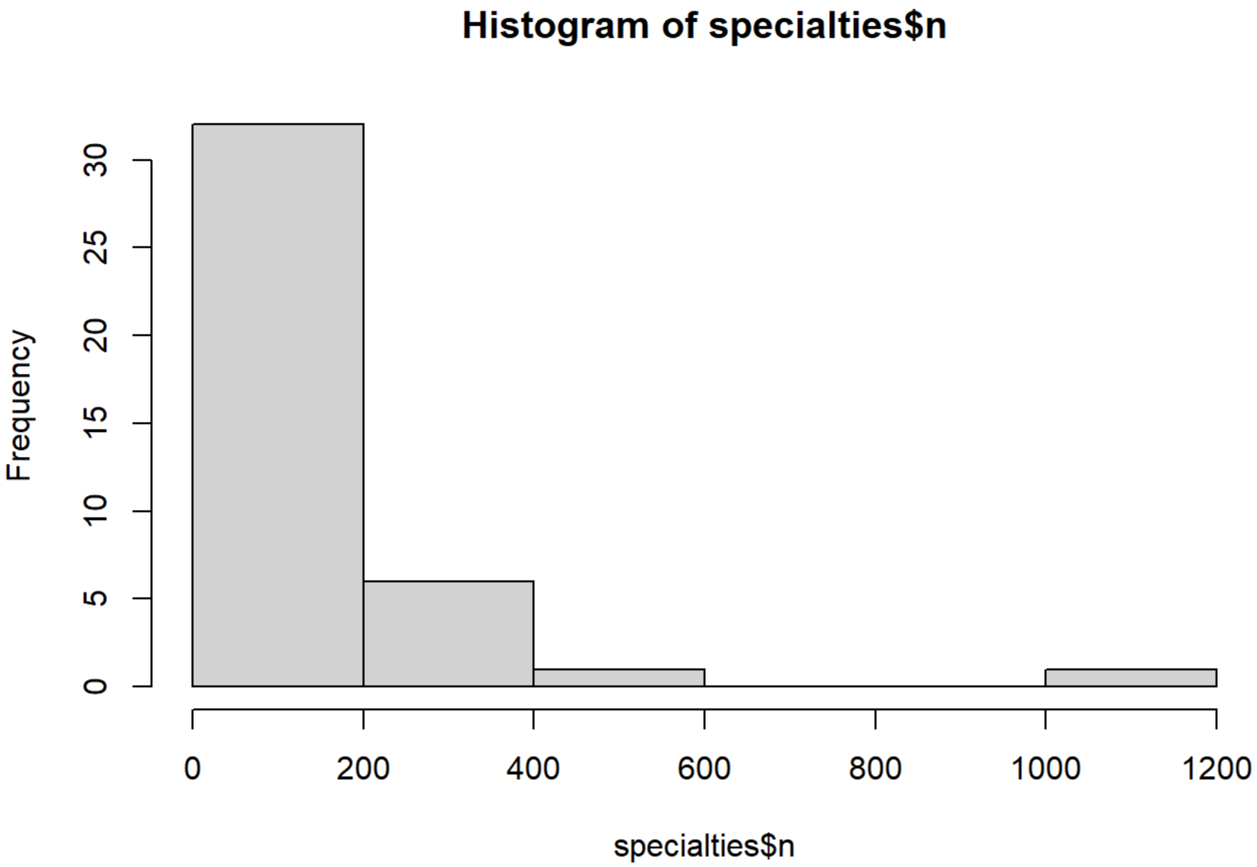
1. Identify specialties

```
specialties <- mtsamples %>%
  count(medical_specialty) %>%
  arrange(desc(n))
specialties
```

```
                medical_specialty    n
1                         Surgery 1103
2      Consult - History and Phy.  516
3      Cardiovascular / Pulmonary  372
4                      Orthopedic  355
5                       Radiology  273
6                General Medicine  259
7                Gastroenterology  230
8                       Neurology  223
9   SOAP / Chart / Progress Notes  166
10        Obstetrics / Gynecology  160
11                        Urology  158
12              Discharge Summary  108
13           ENT - Otolaryngology   98
14                   Neurosurgery   94
15          Hematology - Oncology   90
16                  Ophthalmology   83
17                     Nephrology   81
18         Emergency Room Reports   75
19          Pediatrics - Neonatal   70
20                Pain Management   62
21        Psychiatry / Psychology   53
22                   Office Notes   51
23                       Podiatry   47
24                    Dermatology   29
25     Cosmetic / Plastic Surgery   27
26                      Dentistry   27
27                        Letters   23
28      Physical Medicine - Rehab   21
29                 Sleep Medicine   20
30                  Endocrinology   19
31                     Bariatrics   18
32         IME-QME-Work Comp etc.   16
33                   Chiropractic   14
34           Diets and Nutritions   10
35                   Rheumatology   10
36              Speech - Language    9
37                        Autopsy    8
38       Lab Medicine - Pathology    8
39           Allergy / Immunology    7
40      Hospice - Palliative Care    6
```

From examining the table of medical specialties and their counts, we find that there are 40 specialties in total. The most common specialty by far is “Surgery”, with over 1,100 records. However, there are also more specific specialties that overlap with surgery, such as “Neurosurgery” with 94 records, and “Cosmetic / Plastic Surgery” with only 27 records.

```
hist(specialties$n)
```



From the distribution of the specialty frequencies, we find that they are not evenly distributed; instead, the distribution is heavily right-skewed, with some categories having far more records than others.

2. Tokenize and visualize

```
mtsamples_tokens <- mtsamples %>%
  unnest_tokens(token, transcription)
head(mtsamples_tokens)
```

```
  X                                                     description
1 0  A 23-year-old white female presents with complaint of allergies.
2 0  A 23-year-old white female presents with complaint of allergies.
3 0  A 23-year-old white female presents with complaint of allergies.
4 0  A 23-year-old white female presents with complaint of allergies.
5 0  A 23-year-old white female presents with complaint of allergies.
```

```
6 0  A 23-year-old white female presents with complaint of allergies.
      medical_specialty         sample_name
1  Allergy / Immunology  Allergic Rhinitis
2  Allergy / Immunology  Allergic Rhinitis
3  Allergy / Immunology  Allergic Rhinitis
4  Allergy / Immunology  Allergic Rhinitis
5  Allergy / Immunology  Allergic Rhinitis
6  Allergy / Immunology  Allergic Rhinitis

keywords
1 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
2 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
3 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
4 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
5 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
6 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
       token
1 subjective
2       this
3         23
4       year
5        old
6      white
```

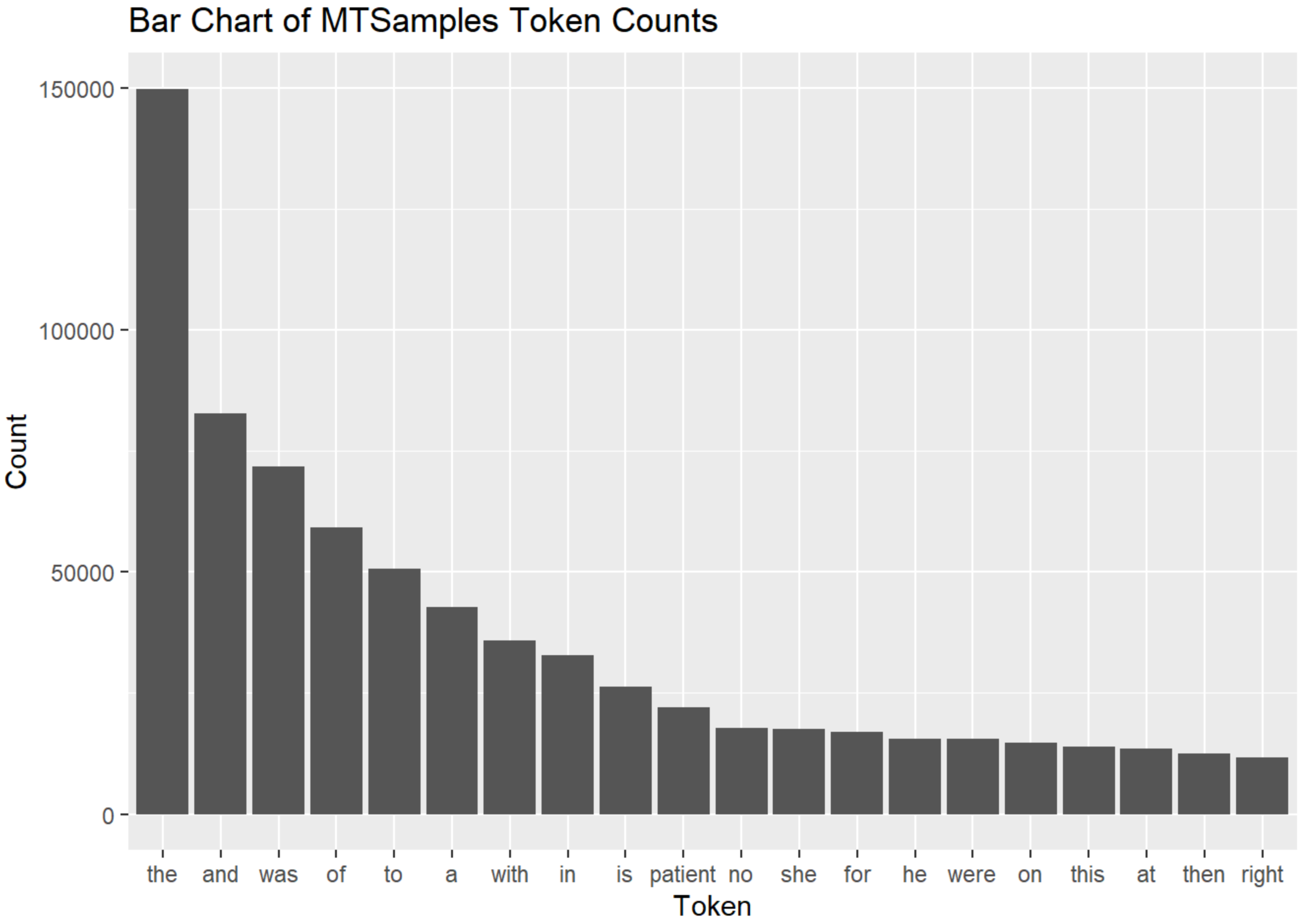
```
head(mtsamples_tokens$token)
```

```
[1] "subjective" "this"       "23"         "year"       "old"
[6] "white"
```

```
mtsamples_token_counts <- mtsamples_tokens %>%
  count(token) %>%
  arrange(desc(n))

head(mtsamples_token_counts)
```

```
  token      n
1   the 149888
2   and  82779
3   was  71765
4    of  59205
5    to  50632
6     a  42812
```

```
ggplot(data = head(mtsamples_token_counts, n = 20), aes(x = fct_reorder(token, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Token Counts",
       y = "Count",
       x = "Token")
```



We see from the bar chart that the most common word in the transcriptions column is “the”, followed by “and”, “was”, “of”, etc. This makes perfect sense because “the” is the most common word in the English language overall, as it is necessary for general communication. Most of these words are stop words that do not provide significant meaning; the exception is “patient”, which is quite common at a count of roughly 20,000, and is expected in a collection of medical descriptions.

3. Remove stop words and visualize

```
mtsamples_tokens <- mtsamples %>%
  unnest_tokens(token, transcription) %>%
  filter(!str_detect(token, "^\\d{1,3}(,\\d{3})*(\\.\\d+)?$")) %>%
  anti_join(stop_words, by = c("token" = "word"))
head(mtsamples_tokens)
```

```
  X                                                    description
1 0  A 23-year-old white female presents with complaint of allergies.
2 0  A 23-year-old white female presents with complaint of allergies.
3 0  A 23-year-old white female presents with complaint of allergies.
4 0  A 23-year-old white female presents with complaint of allergies.
5 0  A 23-year-old white female presents with complaint of allergies.
6 0  A 23-year-old white female presents with complaint of allergies.
      medical_specialty         sample_name
1  Allergy / Immunology  Allergic Rhinitis
2  Allergy / Immunology  Allergic Rhinitis
3  Allergy / Immunology  Allergic Rhinitis
4  Allergy / Immunology  Allergic Rhinitis
5  Allergy / Immunology  Allergic Rhinitis
6  Allergy / Immunology  Allergic Rhinitis

keywords
1 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
2 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
3 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
4 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
5 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
6 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
       token
1 subjective
2      white
3     female
4  complaint
5  allergies
6  allergies
```

```r
head(mtsamples_tokens$token)
```

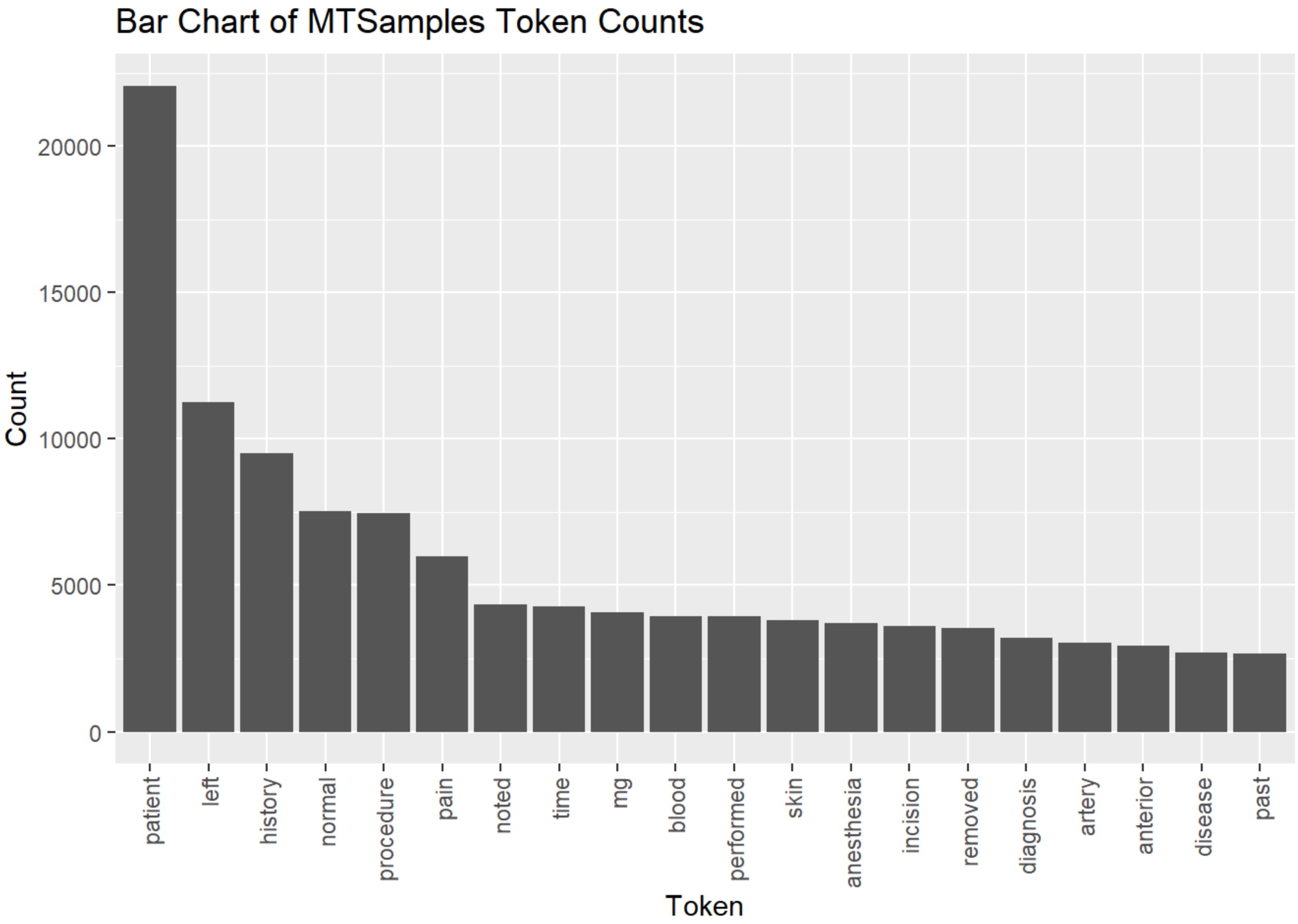
```
[1] "subjective" "white"      "female"     "complaint"  "allergies"
[6] "allergies"
```

```r
mtsamples_token_counts <- mtsamples_tokens %>%
  count(token) %>%
  arrange(desc(n))

head(mtsamples_token_counts)
```

```
      token     n
1   patient 22065
```

```
2       left 11258
3    history  9509
4     normal  7526
5  procedure  7463
6       pain  5976
```

```
ggplot(data = head(mtsamples_token_counts, n = 20), aes(x = fct_reorder(token, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Token Counts",
       y = "Count",
       x = "Token") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



From this bar graph, we get a much more valuable analysis of the most common tokens: We find that “patient” is the most common word that is not a stopword or a number. Some other common tokens are “history”, “procedure”, “pain”, and “time”, which all refer to common measurements in medical cases. Other common tokens refer to specific biological terms like “blood”, “skin”, and “artery”.

4. Tokenize into bigrams and trigrams

Note that this question asks to repeat question 2 (not 3), so we do not remove stopwords or numbers.

We first tokenize into bigrams:

```
mtsamples_tokens <- mtsamples %>%
  unnest_ngrams(token, transcription, n = 2)
head(mtsamples_tokens)
```

```
  X                                                       description
1 0  A 23-year-old white female presents with complaint of allergies.
2 0  A 23-year-old white female presents with complaint of allergies.
3 0  A 23-year-old white female presents with complaint of allergies.
4 0  A 23-year-old white female presents with complaint of allergies.
5 0  A 23-year-old white female presents with complaint of allergies.
6 0  A 23-year-old white female presents with complaint of allergies.
      medical_specialty         sample_name
1  Allergy / Immunology  Allergic Rhinitis
2  Allergy / Immunology  Allergic Rhinitis
3  Allergy / Immunology  Allergic Rhinitis
4  Allergy / Immunology  Allergic Rhinitis
5  Allergy / Immunology  Allergic Rhinitis
6  Allergy / Immunology  Allergic Rhinitis

keywords
1 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
2 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
3 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
4 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
5 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
6 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
             token
1 subjective this
2         this 23
3         23 year
4        year old
5       old white
6    white female
```

```
head(mtsamples_tokens$token)
```

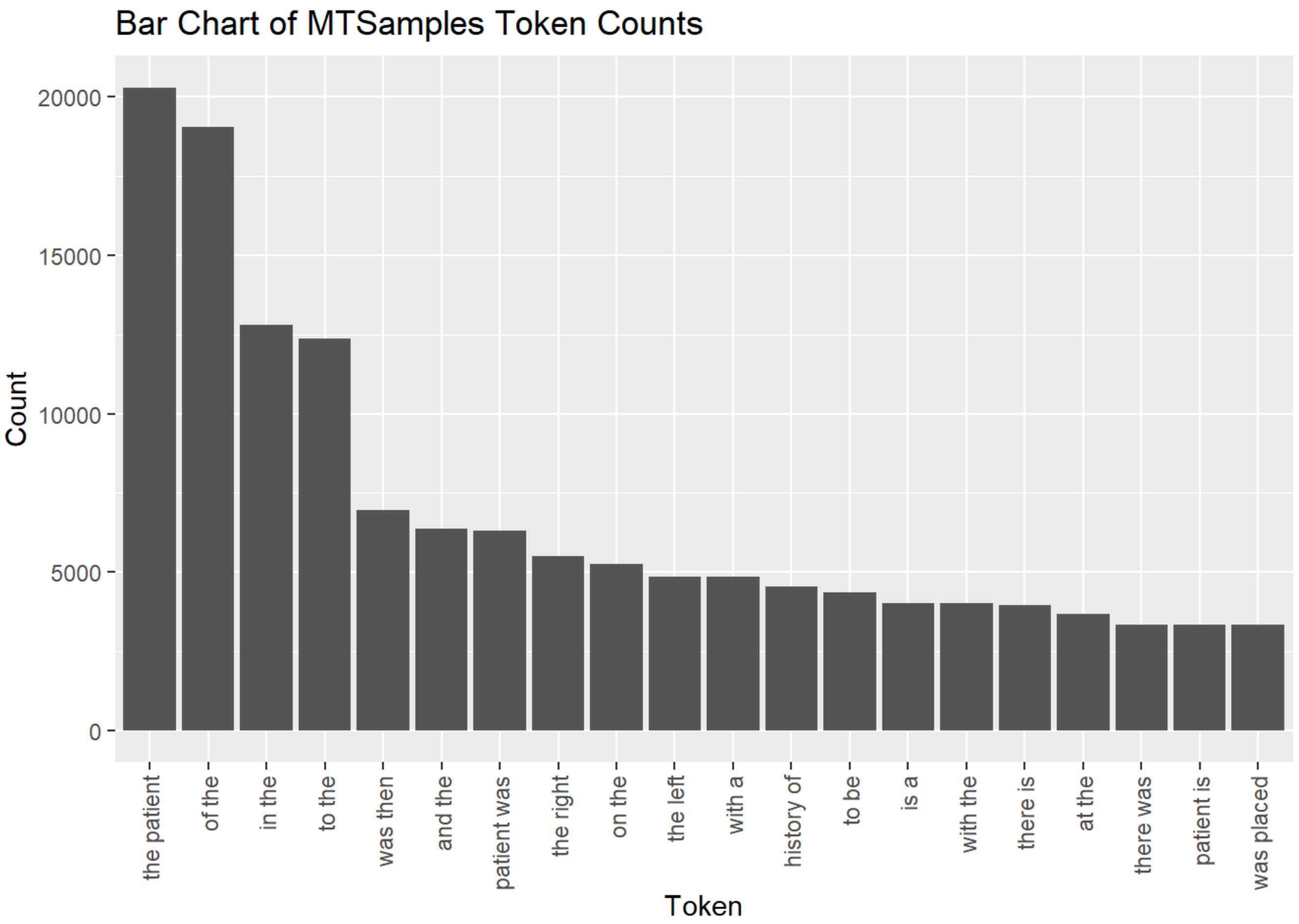
```
[1] "subjective this" "this 23"         "23 year"         "year old"
[5] "old white"       "white female"
```

```
mtsamples_token_counts <- mtsamples_tokens %>%
  count(token) %>%
  arrange(desc(n))
```

```
head(mtsamples_token_counts)
```

```
        token     n
1 the patient 20307
2      of the 19062
3      in the 12790
4      to the 12374
5    was then  6956
6     and the  6350
```

```
ggplot(data = head(mtsamples_token_counts, n = 20), aes(x = fct_reorder(token, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Token Counts",
       y = "Count",
       x = "Token") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



We see that the most common bigram is “the patient” with over 20,000 appearances, which makes perfect sense given that this is the focus of a medical report. Other bigrams also include the patient, such as “patient was” and “patient is”. However, the rest of the bigrams (besides “history of”) do not provide significant information, much like the stopwords from Question 2.

We now tokenize into trigrams:

```r
mtsamples_tokens <- mtsamples %>%
  unnest_ngrams(token, transcription, n = 3)
head(mtsamples_tokens)
```

```
  X                                                         description
1 0  A 23-year-old white female presents with complaint of allergies.
2 0  A 23-year-old white female presents with complaint of allergies.
3 0  A 23-year-old white female presents with complaint of allergies.
4 0  A 23-year-old white female presents with complaint of allergies.
5 0  A 23-year-old white female presents with complaint of allergies.
6 0  A 23-year-old white female presents with complaint of allergies.
      medical_specialty         sample_name
1  Allergy / Immunology  Allergic Rhinitis
2  Allergy / Immunology  Allergic Rhinitis
3  Allergy / Immunology  Allergic Rhinitis
4  Allergy / Immunology  Allergic Rhinitis
5  Allergy / Immunology  Allergic Rhinitis
6  Allergy / Immunology  Allergic Rhinitis

keywords
1 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
2 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
3 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
4 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
5 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
6 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
                  token
1    subjective this 23
2          this 23 year
3           23 year old
4        year old white
5      old white female
6 white female presents
```

```r
head(mtsamples_tokens$token)
```

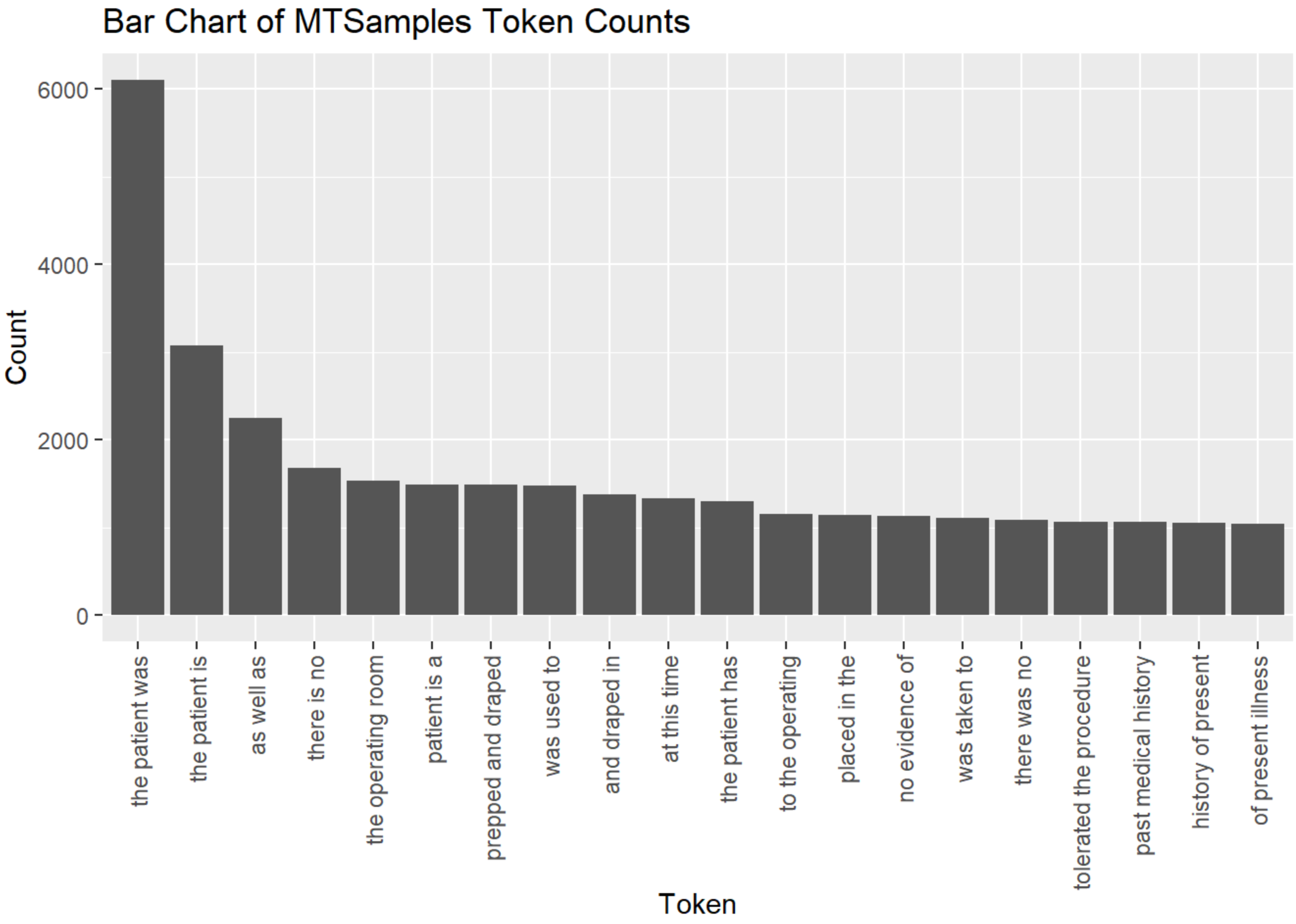
```
[1] "subjective this 23"    "this 23 year"          "23 year old"
[4] "year old white"        "old white female"      "white female presents"
```

```r
mtsamples_token_counts <- mtsamples_tokens %>%
  count(token) %>%
  arrange(desc(n))
```

```
head(mtsamples_token_counts)
```

```
                token    n
1    the patient was 6104
2     the patient is 3075
3         as well as 2243
4        there is no 1678
5 the operating room 1532
6       patient is a 1491
```

```
ggplot(data = head(mtsamples_token_counts, n = 20), aes(x = fct_reorder(token, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Token Counts",
       y = "Count",
       x = "Token") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



The most common trigram was “the patient was”, followed by “the patient is”, which once again makes perfect sense in this context. However, we also see more specific tokens such as “the operating room”, “prepped and draped”, and “tolerated the procedure”, which speaks to the specificity of language used in medical analysis.

5. Count words that appear before and after “patient”

```
patient_tokens <- mtsamples %>%
  unnest_ngrams(ngram, transcription, n = 3) %>%
  separate(ngram, into = c("word1", "word2", "word3"), sep = " ") %>%
  filter(word2 == "patient")
```

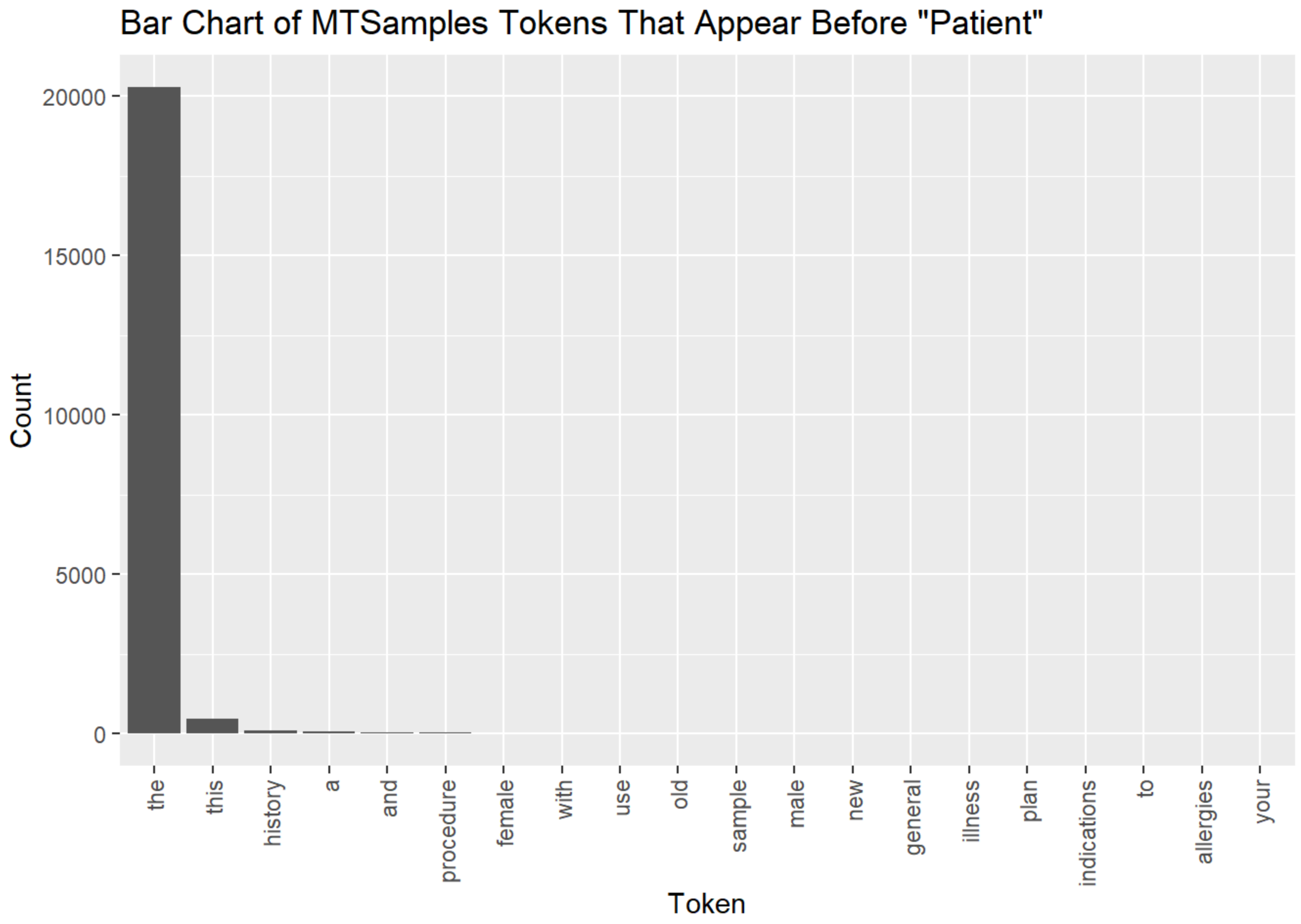
```
head(patient_tokens)
```

```
  X
1 5
2 5
3 5
4 5
5 6
6 6

description
1  Morbid obesity.  Laparoscopic antecolic antegastric Roux-en-Y gastric bypass with EEA
anastomosis.  This is a 30-year-old female, who has been overweight for many years.  She has
tried many different diets, but is unsuccessful.
2  Morbid obesity.  Laparoscopic antecolic antegastric Roux-en-Y gastric bypass with EEA
anastomosis.  This is a 30-year-old female, who has been overweight for many years.  She has
tried many different diets, but is unsuccessful.
3  Morbid obesity.  Laparoscopic antecolic antegastric Roux-en-Y gastric bypass with EEA
anastomosis.  This is a 30-year-old female, who has been overweight for many years.  She has
tried many different diets, but is unsuccessful.
4  Morbid obesity.  Laparoscopic antecolic antegastric Roux-en-Y gastric bypass with EEA
anastomosis.  This is a 30-year-old female, who has been overweight for many years.  She has
tried many different diets, but is unsuccessful.
5                                                                              Liposuction of the
supraumbilical abdomen, revision of right breast reconstruction, excision of soft tissue fullness
of the lateral abdomen and flank.\n
6                                                                              Liposuction of the
supraumbilical abdomen, revision of right breast reconstruction, excision of soft tissue fullness
of the lateral abdomen and flank.\n
  medical_specialty                   sample_name
1        Bariatrics  Laparoscopic Gastric Bypass
2        Bariatrics  Laparoscopic Gastric Bypass
3        Bariatrics  Laparoscopic Gastric Bypass
4        Bariatrics  Laparoscopic Gastric Bypass
5        Bariatrics                  Liposuction
6        Bariatrics                  Liposuction

keywords
1                                                  bariatrics, gastric bypass, eea anastomosis,
roux-en-y, antegastric, antecolic, morbid obesity, roux limb, gastric pouch, intubation,
laparoscopic, bypass, roux, endotracheal, anastomosis, gastric
2                                                  bariatrics, gastric bypass, eea anastomosis,
roux-en-y, antegastric, antecolic, morbid obesity, roux limb, gastric pouch, intubation,
laparoscopic, bypass, roux, endotracheal, anastomosis, gastric
3                                                  bariatrics, gastric bypass, eea anastomosis,
roux-en-y, antegastric, antecolic, morbid obesity, roux limb, gastric pouch, intubation,
laparoscopic, bypass, roux, endotracheal, anastomosis, gastric
4                                                  bariatrics, gastric bypass, eea anastomosis,
roux-en-y, antegastric, antecolic, morbid obesity, roux limb, gastric pouch, intubation,
laparoscopic, bypass, roux, endotracheal, anastomosis, gastric
5 bariatrics, breast reconstruction, excess, lma anesthesia, lipodystrophy, liposuction, abdomen,
drain site, flank, latissimus dorsi flap, soft tissue, supraumbilical, surgical bra,
supraumbilical abdomen, reconstruction, breast, tissue, implant,
6 bariatrics, breast reconstruction, excess, lma anesthesia, lipodystrophy, liposuction, abdomen,
drain site, flank, latissimus dorsi flap, soft tissue, supraumbilical, surgical bra,
supraumbilical abdomen, reconstruction, breast, tissue, implant,
  word1   word2       word3
1   the patient   procedure
2   the patient         was
3   the patient          in
4   the patient   tolerated
5   the patient          is
6   the patient significant
```

```r
patient_token_counts <- patient_tokens %>%
  count(word1, sort = TRUE)
head(patient_token_counts)
```

```
      word1     n
1       the 20294
2      this   463
3   history   101
4         a    67
5       and    47
6 procedure    32
```

```r
ggplot(data = head(patient_token_counts, n = 20), aes(x = fct_reorder(word1, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Tokens That Appear Before \"Patient\"",
       y = "Count",
       x = "Token") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

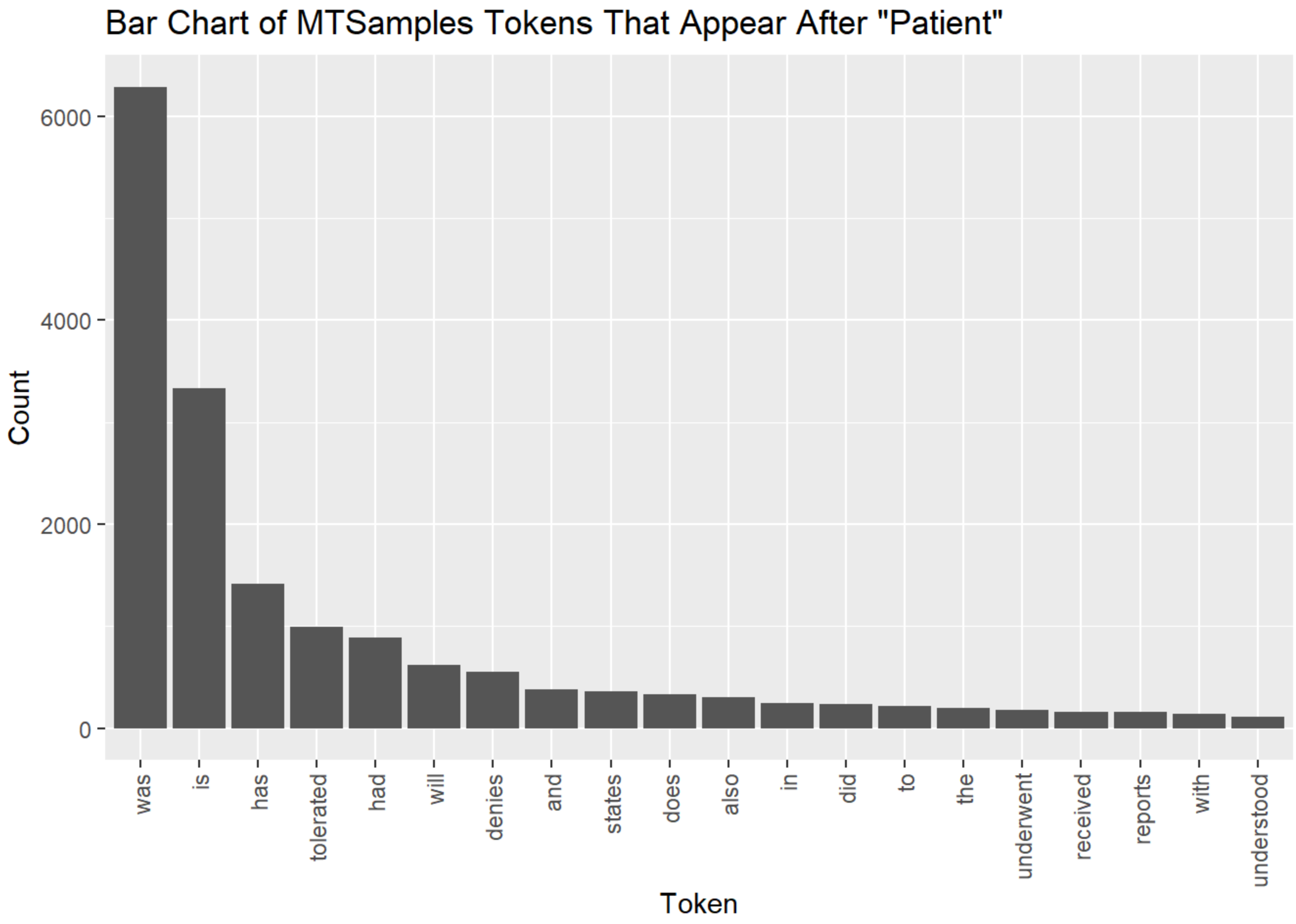


From the bar chart, we find that the most common word that appears before “patient” is overwhelmingly “the”, followed by “this”, with the other tokens being extremely obscure by comparison. Again, this makes sense, since the word “patient” is a noun and is typically followed by either “the” or “this”.

```r
patient_token_counts <- patient_tokens %>%
  count(word3, sort = TRUE)
head(patient_token_counts)
```

```
      word3    n
1       was 6291
2        is 3332
3       has 1417
4 tolerated  994
5       had  886
6      will  616
```

```r
ggplot(data = head(patient_token_counts, n = 20), aes(x = fct_reorder(word3, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Tokens That Appear After \"Patient\"",
       y = "Count",
       x = "Token") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



We find that the distribution of words that appear after “patient” is much more evenly distributed; however, the most common word after “patient” is “was”, followed by “is”, “has”, and “tolerated”. It makes sense that these words are either stopwords or verbs in the past tense, since the typical format of a medical report is to describe what happened to a patient in past tense.

6. Find most frequent words in each speciality

```
mtsamples_tokens <- mtsamples %>%
  unnest_tokens(token, transcription) %>%
  filter(!str_detect(token, "^\\d{1,3}(,\\d{3})*(\\.\\d+)?$")) %>%
  anti_join(stop_words, by = c("token" = "word"))
```

```
mtsamples_token_counts <- mtsamples_tokens %>%
  count(medical_specialty, token, sort = TRUE) %>%
  group_by(medical_specialty) %>%
  slice_max(n, n = 5, with_ties = FALSE) %>%
  ungroup()

mtsamples_token_counts
```

```
# A tibble: 200 × 3
   medical_specialty       token         n
   <chr>                   <chr>     <int>
 1 " Allergy / Immunology" history      38
 2 " Allergy / Immunology" noted        23
 3 " Allergy / Immunology" patient      22
 4 " Allergy / Immunology" allergies    21
 5 " Allergy / Immunology" nasal        13
 6 " Autopsy"              left         83
 7 " Autopsy"              inch         59
 8 " Autopsy"              neck         55
 9 " Autopsy"              anterior     47
10 " Autopsy"              body         40
# ℹ 190 more rows
```

7. Identify interesting 5-grams

```r
mtsamples_tokens <- mtsamples %>%
  unnest_ngrams(token, transcription, n = 5)
head(mtsamples_tokens)
```

```
  X                                                        description
1 0  A 23-year-old white female presents with complaint of allergies.
2 0  A 23-year-old white female presents with complaint of allergies.
3 0  A 23-year-old white female presents with complaint of allergies.
4 0  A 23-year-old white female presents with complaint of allergies.
5 0  A 23-year-old white female presents with complaint of allergies.
6 0  A 23-year-old white female presents with complaint of allergies.
       medical_specialty         sample_name
1  Allergy / Immunology  Allergic Rhinitis
2  Allergy / Immunology  Allergic Rhinitis
3  Allergy / Immunology  Allergic Rhinitis
4  Allergy / Immunology  Allergic Rhinitis
5  Allergy / Immunology  Allergic Rhinitis
6  Allergy / Immunology  Allergic Rhinitis

keywords
1 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
2 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
3 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
4 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
5 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
6 allergy / immunology, allergic rhinitis, allergies, asthma, nasal sprays, rhinitis, nasal,
erythematous, allegra, sprays, allergic,
                                   token
1           subjective this 23 year old
```

```
2                this 23 year old white
3             23 year old white female
4       year old white female presents
5       old white female presents with
6 white female presents with complaint
```

```
head(mtsamples_tokens$token)
```

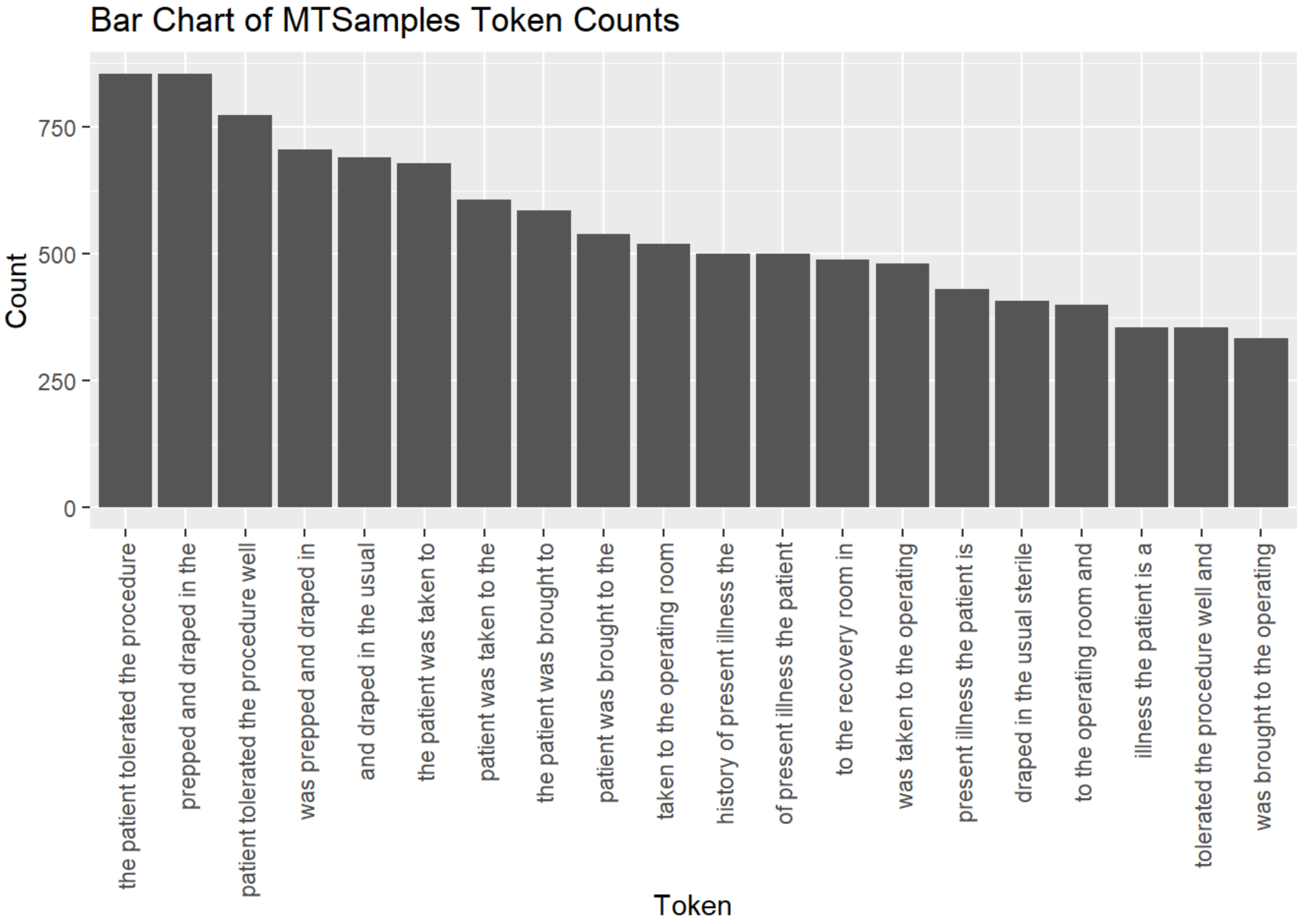
```
[1] "subjective this 23 year old"
[2] "this 23 year old white"
[3] "23 year old white female"
[4] "year old white female presents"
[5] "old white female presents with"
[6] "white female presents with complaint"
```

```
mtsamples_token_counts <- mtsamples_tokens %>%
  count(token) %>%
  arrange(desc(n))
```

```
head(mtsamples_token_counts)
```

```
                                 token   n
1  the patient tolerated the procedure 855
2            prepped and draped in the 854
3 patient tolerated the procedure well 773
4            was prepped and draped in 706
5              and draped in the usual 690
6             the patient was taken to 678
```

```
ggplot(data = head(mtsamples_token_counts, n = 20), aes(x = fct_reorder(token, -n), y = n)) +
  geom_bar(stat = "identity") +
  labs(title = "Bar Chart of MTSamples Token Counts",
       y = "Count",
       x = "Token") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```



Interestingly, nearly all of the most common 5-grams (groups of 5 words) in the transcription data describe either the patient tolerating a procedure or the patient being draped in sterile robes and taken to the operating room, which again highlights the necessity of consistent language between medical reports.