In [1]:

```
import re
import numpy as np
import pandas as pd

import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from wordcloud import WordCloud, STOPWORDS

import nltk
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
stopword = set(stopwords.words('english'))
```

In [2]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [3]:

```
df = pd.read_csv('covid_abstracts.csv')
```

In [4]:

```
df.head()
```

Out[4]:

	title	abstract	url
0	Real-World Experience with COVID-19 Including...	This article summarizes the experiences of COV...	https://pubmed.ncbi.nlm.nih.gov/35008137
1	Successful outcome of pre-engraftment COVID-19...	Coronavirus disease 2019 COVID-19 caused by...	https://pubmed.ncbi.nlm.nih.gov/35008104
2	The impact of COVID-19 on oncology professiona...	BACKGROUND COVID-19 has had a significant imp...	https://pubmed.ncbi.nlm.nih.gov/35007996
3	ICU admission and mortality classifiers for CO...	The coronavirus disease 2019 COVID-19 which ...	https://pubmed.ncbi.nlm.nih.gov/35007991
4	Clinical evaluation of nasopharyngeal midturb...	In the setting of supply chain shortages of na...	https://pubmed.ncbi.nlm.nih.gov/35007959

In [5]:

```
df.tail()
```

Out[5]:

	title	abstract	url
9995	Rooming-in Breastfeeding and Neonatal Follow-...	INTRODUCTION Due to growing evidence suggesti...	https://pubmed.ncbi.nlm.nih.gov/34851815
9996	Acute Retinal Necrosis from Reactivation of Va...	PURPOSE To report a case of acute retinal nec...	https://pubmed.ncbi.nlm.nih.gov/34851795
9997	Acute Abducens Nerve Palsy Following the Secon...	The authors report the case of an otherwise he...	https://pubmed.ncbi.nlm.nih.gov/34851785
9998	Planning and Implementing the Protocol for Psy...	The present study aims to plan the protocol fo...	https://pubmed.ncbi.nlm.nih.gov/34851781
9999	Prolonged corrected QT interval in hospitalize...	OBJECTIVE To evaluate the association of a pr...	https://pubmed.ncbi.nlm.nih.gov/34851769

In [6]:

```
df.shape
```

Out[6]:

```
(10000, 3)
```

In [7]:

```
df.columns
```

Out[7]:

```
Index(['title', 'abstract', 'url'], dtype='object')
```

In [8]:

```
df.duplicated().sum()
```

Out[8]:

```
0
```

In [9]:

```
df.isnull().sum()
```

Out[9]:

```
title       0
abstract    0
url         0
dtype: int64
```

In [10]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   title     10000 non-null  object
 1   abstract  10000 non-null  object
 2   url       10000 non-null  object
dtypes: object(3)
memory usage: 234.5+ KB
```

In [11]:

```
df.nunique()
```

Out[11]:

```
title       10000
abstract    10000
url         10000
dtype: int64
```

In [12]:

```
def remove_punctuation(text):

    # punctuations except -
    punc ='''?!.,:;_–[](){}'"`~|\/@#$%^&+=*'''
    for i in text:
        if i in punc:
            text = text.replace(i, ' ')
    return text.strip()


def word_count(text):

    # word tokenization
    lst = word_tokenize(text)
    return len(lst)


def preprocess(text):

    # lower casing
    text=text.lower()

    # stopword removal
    text = [word for word in text.split(' ') if word not in stopword]
    text=" ".join(text)

    # lemmatization
    text = [lemmatizer.lemmatize(word) for word in text.split(' ')]
    text = " ".join(text)

    # remove extra spaces
    text = re.sub("\s\s+", " ", text)
    return text.strip()
```

In [13]:

```
# apply functions
df['title']=df['title'].apply(remove_punctuation)
df['wc_title']=df['title'].apply(word_count)
```

In [14]:

```
df
```

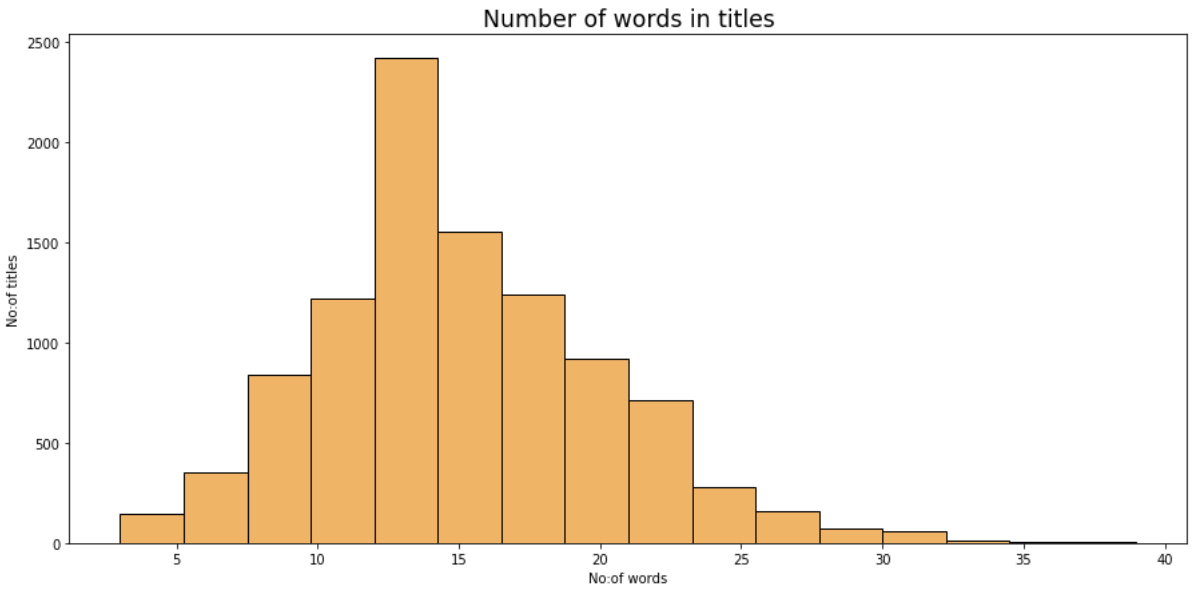
Out[14]:

	title	abstract	url	wc_title
0	Real-World Experience with COVID-19 Including...	This article summarizes the experiences of COV...	https://pubmed.ncbi.nlm.nih.gov/35008137	21
1	Successful outcome of pre-engraftment COVID-19...	Coronavirus disease 2019 COVID-19 caused by...	https://pubmed.ncbi.nlm.nih.gov/35008104	16
2	The impact of COVID-19 on oncology professiona...	BACKGROUND COVID-19 has had a significant imp...	https://pubmed.ncbi.nlm.nih.gov/35007996	19
3	ICU admission and mortality classifiers for CO...	The coronavirus disease 2019 COVID-19 which ...	https://pubmed.ncbi.nlm.nih.gov/35007991	18
4	Clinical evaluation of nasopharyngeal midturb...	In the setting of supply chain shortages of na...	https://pubmed.ncbi.nlm.nih.gov/35007959	14
...	...	...	...	...
9995	Rooming-in Breastfeeding and Neonatal Follow-...	INTRODUCTION Due to growing evidence suggesti...	https://pubmed.ncbi.nlm.nih.gov/34851815	12
9996	Acute Retinal Necrosis from Reactivation of Va...	PURPOSE To report a case of acute retinal nec...	https://pubmed.ncbi.nlm.nih.gov/34851795	14
9997	Acute Abducens Nerve Palsy Following the Secon...	The authors report the case of an otherwise he...	https://pubmed.ncbi.nlm.nih.gov/34851785	13
9998	Planning and Implementing the Protocol for Psy...	The present study aims to plan the protocol fo...	https://pubmed.ncbi.nlm.nih.gov/34851781	17
9999	Prolonged corrected QT interval in hospitalize...	OBJECTIVE To evaluate the association of a pr...	https://pubmed.ncbi.nlm.nih.gov/34851769	20

10000 rows × 4 columns

In [15]:

```
# plot
plt.figure(figsize=(15,7))
ax=sns.histplot(x='wc_title', data=df, bins=16, color='#eb9b34')
plt.title('Number of words in titles',size='xx-large')
plt.xlabel('No:of words')
plt.ylabel('No:of titles')
plt.show()
```

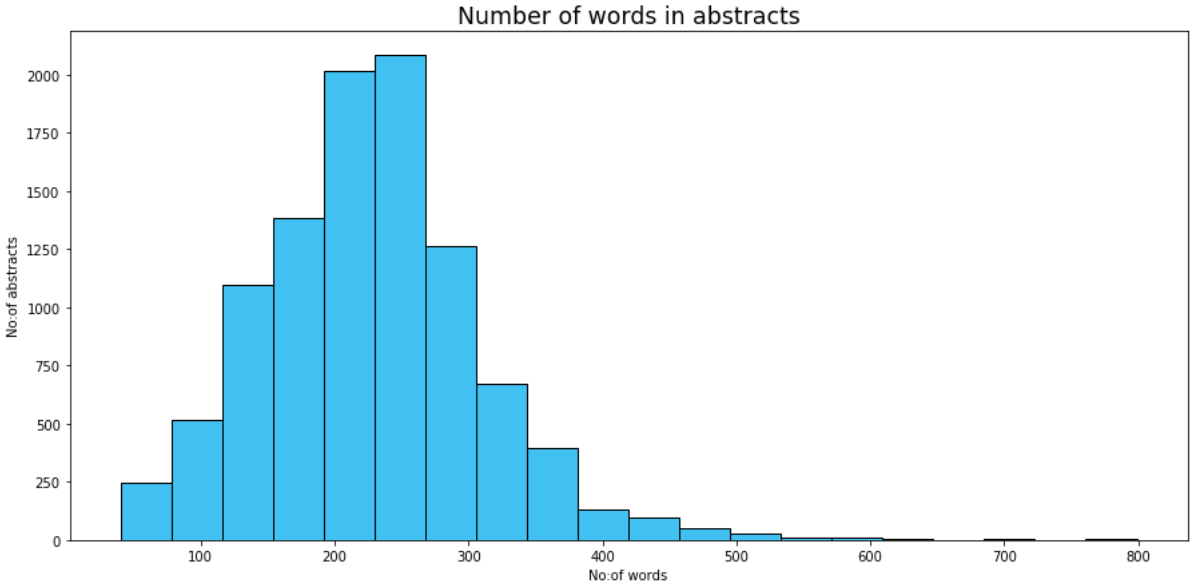


In [16]:

```
# apply functions
df['abstract']=df['abstract'].apply(remove_punctuation)
df['wc_abstract']=df['abstract'].apply(word_count)
```

In [17]:

```
# plot
plt.figure(figsize=(15,7))
ax=sns.histplot(x='wc_abstract', data=df, bins=20, color='#00acee')
plt.title('Number of words in abstracts',size='xx-large')
plt.xlabel('No:of words')
plt.ylabel('No:of abstracts')
plt.show()
```

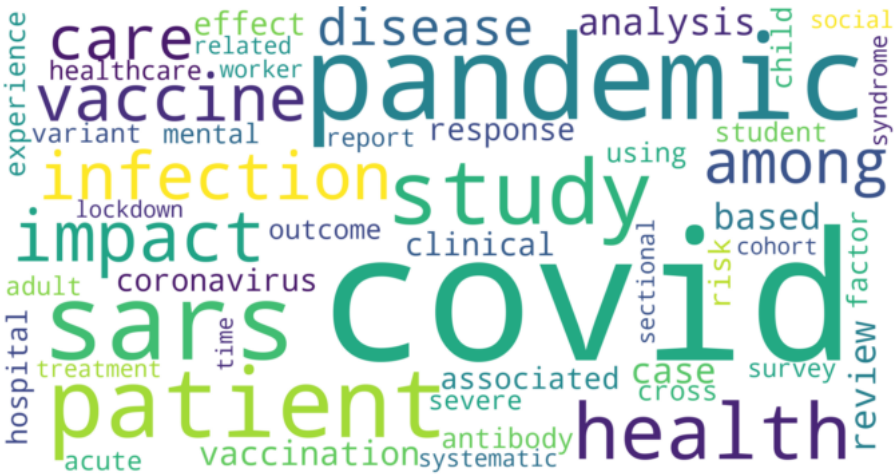


In [18]:

```
# preprocess text
df['cleaned_title']=df['title'].apply(preprocess)
```

In [19]:

```
title = ' '.join(df['cleaned_title'])

# generate Word Cloud
word_cloud = WordCloud(collocations=False,
                       background_color='white',max_words=50, stopwords=STOPWORDS, min_word_length=4,
                       width=2048, height=1080).generate(title)

# Display the generated Word Cloud
plt.figure(figsize=(12,8))
plt.imshow(word_cloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

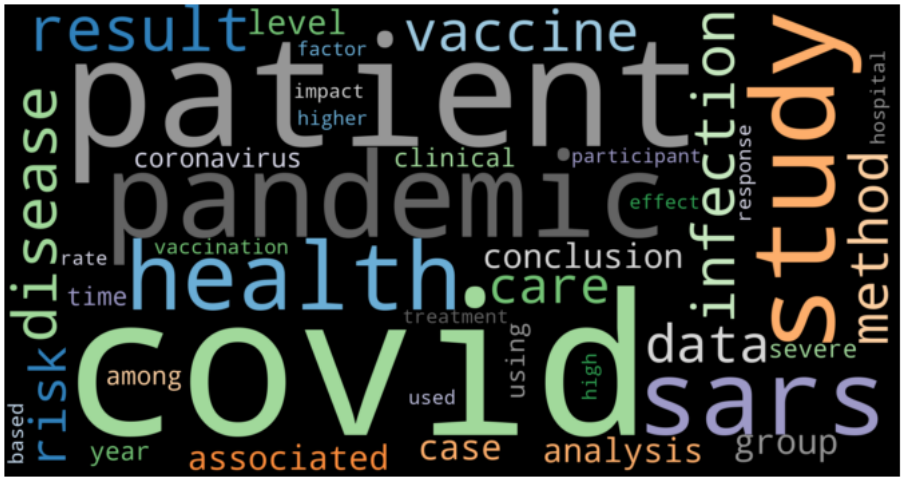


In [20]:

```
# preprocess text
df['cleaned_abstract']=df['abstract'].apply(preprocess)
```

In [21]:

```
abstracts = ' '.join(df['cleaned_abstract'])

# generate Word Cloud
word_cloud = WordCloud(collocations=False,
                       background_color='black',max_words=40, stopwords=STOPWORDS, min_word_length=4,
                       colormap='tab20c',width=2048, height=1080).generate(abstracts)

# Display the generated Word Cloud
plt.figure(figsize=(12,8))
plt.imshow(word_cloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



In [22]:

```
df
```

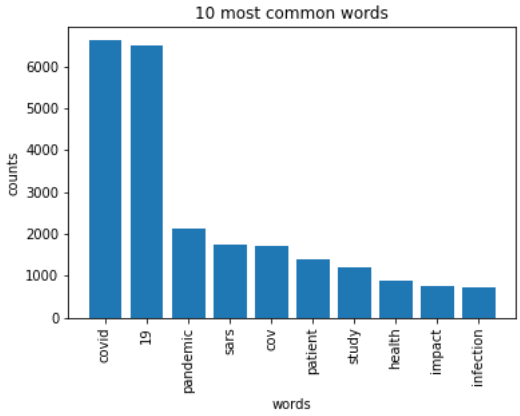
Out[22]:

	title	abstract	url	wc_title	wc_abstract	cleaned_title	cleaned_abstract
0	Real-World Experience with COVID-19 Including...	This article summarizes the experiences of COV...	https://pubmed.ncbi.nlm.nih.gov/35008137	21	264	real-world experience covid-19 including direc...	article summarizes experience covid-19 patient...
1	Successful outcome of pre-engraftment COVID-19...	Coronavirus disease 2019 COVID-19 caused by...	https://pubmed.ncbi.nlm.nih.gov/35008104	16	200	successful outcome pre-engraftment covid-19 hc...	coronavirus disease 2019 covid-19 caused sever...
2	The impact of COVID-19 on oncology professiona...	BACKGROUND COVID-19 has had a significant imp...	https://pubmed.ncbi.nlm.nih.gov/35007996	19	315	impact covid-19 oncology professionals-one yea...	background covid-19 significant impact well-be...
3	ICU admission and mortality classifiers for CO...	The coronavirus disease 2019 COVID-19 which ...	https://pubmed.ncbi.nlm.nih.gov/35007991	18	299	icu admission mortality classifier covid-19 pa...	coronavirus disease 2019 covid-19 caused sever...
4	Clinical evaluation of nasopharyngeal midturb...	In the setting of supply chain shortages of na...	https://pubmed.ncbi.nlm.nih.gov/35007959	14	164	clinical evaluation nasopharyngeal midturbinat...	setting supply chain shortage nasopharyngeal n...
...	...	...	...	...	...	...	...
9995	Rooming-in Breastfeeding and Neonatal Follow-...	INTRODUCTION Due to growing evidence suggesti...	https://pubmed.ncbi.nlm.nih.gov/34851815	12	305	rooming-in breastfeeding neonatal follow-up in...	introduction due growing evidence suggesting c...
9996	Acute Retinal Necrosis from Reactivation of Va...	PURPOSE To report a case of acute retinal nec...	https://pubmed.ncbi.nlm.nih.gov/34851795	14	129	acute retinal necrosis reactivation varicella ...	purpose report case acute retinal necrosis arn...
9997	Acute Abducens Nerve Palsy Following the Secon...	The authors report the case of an otherwise he...	https://pubmed.ncbi.nlm.nih.gov/34851785	13	105	acute abducens nerve palsy following second do...	author report case otherwise healthy 65-year-o...
9998	Planning and Implementing the Protocol for Psy...	The present study aims to plan the protocol fo...	https://pubmed.ncbi.nlm.nih.gov/34851781	17	196	planning implementing protocol psychosocial in...	present study aim plan protocol providing psyc...
9999	Prolonged corrected QT interval in hospitalize...	OBJECTIVE To evaluate the association of a pr...	https://pubmed.ncbi.nlm.nih.gov/34851769	20	199	prolonged corrected qt interval hospitalized p...	objective evaluate association prolonged corre...

10000 rows × 7 columns

In [28]:

```
from sklearn.feature_extraction.text import CountVectorizer

# Helper function
def plot_10_most_common_words(count_data, count_vectorizer):
    import matplotlib.pyplot as plt
    words = count_vectorizer.get_feature_names()
    total_counts = np.zeros(len(words))
    for t in count_data:
        total_counts+=t.toarray()[0]

    count_dict = (zip(words, total_counts))
    count_dict = sorted(count_dict, key=lambda x:x[1], reverse=True)[0:10]
    words = [w[0] for w in count_dict]
    counts = [w[1] for w in count_dict]
    x_pos = np.arange(len(words))

    plt.bar(x_pos, counts,align='center')
    plt.xticks(x_pos, words, rotation=90)
    plt.xlabel('words')
    plt.ylabel('counts')
    plt.title('10 most common words')
    plt.show()

# Initialise the count vectorizer with the English stop words
count_vectorizer = CountVectorizer(stop_words ='english')

# Fit and transform the processed titles
count_data = count_vectorizer.fit_transform(df['cleaned_title'])

# Visualise the 10 most common words

plot_10_most_common_words(count_data, count_vectorizer)
```

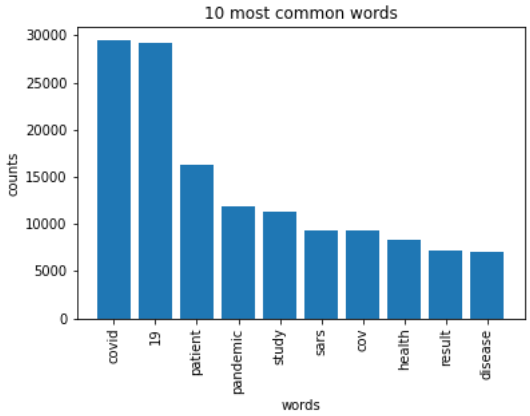


In [29]:

```
count_data_abs = count_vectorizer.fit_transform(df['cleaned_abstract'])
```

In [30]:

```
plot_10_most_common_words(count_data_abs, count_vectorizer)
```



In [31]:

```
from sklearn.decomposition import LatentDirichletAllocation as LDA

# Helper function
def print_topics(model, count_vectorizer, n_top_words):
    words = count_vectorizer.get_feature_names()
    for topic_idx, topic in enumerate(model.components_):
        print("\nTopic #%d:" % topic_idx)
        print(" ".join([words[i]
                        for i in topic.argsort()[:-n_top_words - 1:-1]]))

# Tweak the two parameters below (use int values below 15)
number_topics = 10
number_words = 10

# Create and fit the LDA model
lda = LDA(n_components=number_topics)
lda.fit(count_data)

# Print the topics found by the LDA model
print("Topics found via LDA:")
print_topics(lda, count_vectorizer, number_words)
```

```
Topics found via LDA:

Topic #0:
cl2 496 criticized blackfirst csrc 3x 6172 crhr bedding alzheimer

Topic #1:
cl2 496 ante 14 catarrhal 4750 1976099 chewing commencing 6654

Topic #2:
4m 00244 brackish compressible 515 compressibility climax 27186 4750 androgenic

Topic #3:
4m 00244 british chloroquine compressible 13108567 brackish 3373 bona congregation

Topic #4:
4m 00244 brackish 2793 chaudhry clause clip 439 canceled cvds

Topic #5:
2972 corresponded crhr centrifugation analyzing 00244 4m anosmic ante chewing

Topic #6:
crhr 00244 4m crgns compressible 0064 battery 0068 albania cvi

Topic #7:
4m 00244 british 7274 brackish confuse anthropogenic 2793 ante asw

Topic #8:
4m 00244 brackish albacete 27186 autoptic androgenic 30b compressible albania

Topic #9:
4m 00244 574 comparedimmune artisan cq brackish creator 1976099 addressing
```