(ii)

Answer to worded section: There are two cities (Fisher's Island and Alta) for which

the residents could easily be identified.

These are the only two cities which have zip codes that do not match the rest of

the state they are a part of (in the first three digits), and so if a resident of these

cities are part of a record containing both their state and three-digit zip code then they

can be easily identified as living in their respective cities. Combined with sex age and race

these individuals are very vulnerable to re-identification.

(iii)

One of state or three-digit zip code should be omitted from the dataset. It is more informative

to include the three-digit zip code, it gives more specific information about a patient's location,

and on its own is not unsafe to provide. Therefore, the three-digit zip code should remain in the

dataset and the state should be removed to maximise the dataset's value while minimising risk of

re-identification.

part (E (ii)) written answer

Pros: -Allows the data to be used in some capacity rather than discarded.

-Protects the participants from re-identification by removing the specific

location information.

Cons: -Lose the locational information of the patients put into group 000.

-If there are only a few 'small' zip codes then it is still potentially possible

to re-identify the participants using additional external datasets.

part (E (iii)) written answer

Algorithmically merge all the patients belonging to 'small' zip codes into neighbouring zip

codes. This should be done so that ideally their state doesn't change. This will preserve all

of their data only smearing their location slightly. It also completely removes the risk of

re-identification.

part (E (iv)) answer

Calculating the population density in people per square mile