1

I can think of a few reasons (in no order):

> **LLN**: Doctors with less patients will see higher variance in average outcome, since the LLN hasn't "kicked in" yet. Along the same line of thought, I would expect that doctors at the bottom of the list also have a small number of patients.

Assuming that this dataset may also include doctors who are no longer practicing:

> **Malpractice**: Incompetent doctors who incorrectly treat their patients will make their patients more likely to suffer a sudden cardiac death. These doctors will be more likely to be delicensed and/or be sued for malpractice, which would presumably make them more likely to quit. This would create a correlation between number of patients a doctor has had and the proportion of SCD among those patients.
>
> **Quitting doctors**: Regardless of quality, some doctors are bound to have "lucky" patients and others are bound to have "unlucky" ones. Doctors with unlucky patients who suffer SCD are presumably more likely to quit (they are human, after all).

2

For:

> **Non-linearity**: There may be non-linear effects at play here. If SCD is not a monotonic or continuous function of EF, keeping EF as a continuous variable could end up providing a mis-representative result. It may be, for example, that being in the “high” and “healthy” range have little effect on SCD, and that only being in the “low” range makes a significant difference to SCD outcomes. We won't be able to distinguish any such piecewise effect if we keep the EF variable continuous, but we will be able to see these effects if we categorize EF.
>
> **Ease of Interpretation**: Regression coefficients of “low”, “healthy”, and “high” will be easily interpretable, especially by doctors (who may not have a strong grasp of statistics) who would like to explain risks to patients.

Against:

> **Loss of information**: Information is inherently lost when converting a continuous variable to categorical. This may not end up causing any problems, but we cannot be certain.
>
> **Bunching**: If this is a commonly used classification scheme, doctors may be incentivized to classify patients as one type or another. For example, insurance payouts may differ according to the same low/healthy/high classification scheme that we use. If so, this may incentivize upcoding or introduce other measurement error which would affect the categorical variable more severely than the continuous.

Arbitrary/Mis-categorization: As an example, it may be the case that a “healthy” EF range is actually 50-60, and individuals in this range are very unlikely to suffer SCD. In this case, our model (using a range of 50-65) would over-estimate the probability that a patient with a “healthy” EF suffers SCD, and under-estimate the probability that a patient with a “low” EF suffers SCD.

3

It depends on how we define "performance." For example, it will almost certainly increase the regression model's un-adjusted R2 value. However, it is unclear (at least, to me) what unobserved covariates this variable has, so it may actually become more difficult to interpret the model's parameters; from this more holistic lens, it is less clear that including this variable would improve performance.

Despite my devil's advocacy here, in general, I would still guess that this variable would improve our model's performance because it contains relevant information that is not present in the model's current variables. For example, this variable brings information describing some combination of each patient's 1) willingness to visit the doctor; 2) commitment to their own health; and 3) ability to afford doctor's visits (whether paid out-of-pocket or through insurance, this is almost certainly correlated with income). Each of these three variables is surely related to SCD, so their (indirect) inclusion will likely improve the model's performance.

4

I would not include this variable until I have a better qualitative and holistic understanding of the underlying physiological, bureaucratic, and other factors that formed this dataset. For example, it is possible that patients over a certain age and over a certain EF are recommended to visit their doctor every month, in which case this variable would have very little predictive power.

That being said, the purpose of this project is to "estimate patient risk of suffering a ‘Sudden Cardiac Death’." If we are not interested in understanding the marginal effects of each variable, and including the "patient visit count" variable improves our model's predictive power (in whatever metric(s) we deem relevant), then we may wish to include it anyway.

5

Splitting by "visit ID" rather than by "patient ID". If, as I posit above, the number of doctor’s visits is not independent from the severity of a patient's illness, then sicker patients (who therefore have had more visits) will be over-represented in the RF model's training (note that this is the case regardless of whether "patient visit count" is included).

I would fix this training process by taking a 50-50 split of *patient* IDs to create training and testing data (although 50% is a more even split than I've usually seen...).

6

This prediction does seem high, although I can think of a few reasons that it would still be valid.

> 1) This comparison depends on (among other things) how a 'healthy' EF is defined. Is it healthy relative to other patients in the dataset, some of whom have extremely high EFs? Or is it healthy on an absolute, physiological scale. If it is healthy relative to other patients in the dataset, then there may not be any contradiction in this 1-to-10% comparison.
>
> 2) Data collection and accuracy may contribute to this discrepancy. It's possible that patients in this dataset who have had their EF measured are known to be more at-risk for SCD. If so, if they experience SCD, it is plausibly more likely that their death will be correctly attributed to SCD. In contrast, it is possible that members of the general population experience SCD at similar rates but it is less likely to be detected in their autopsies.
>
> 3) Population differences (selection bias / conditional probability) may also be at play here. This dataset covers US adults who have been to the doctor (in some cases, many times) and who therefore may be less healthy than the US adult population. In other words, our model is telling us that, conditional on being part of this dataset, a 'healthy' 80-year old is estimated to have a 10% risk of sudden cardiac death in the next year. This does not necessarily contradict the statement that "less than 1% of US adults over the age of 80 [die of SCD each year]." That being said, my gut tells me that this disparity is too high to only be explained by conditional probability.

These examples aside, it is entirely possible (and probably likely) that the model's prediction is higher than the real likelihood. For example, there could be a small sample size at the margins of the age distributions. This question does not tell me how many observations there are of patients age 80+, and if there are too few, the predicted likelihood of a 'healthy' 80-year old's SCD may have a very large confidence interval (to this point, the prediction's confidence interval is not mentioned in the question).