CIND820 – Literature Review

Benjamin Faulknor

Table of Contents

[Topic Introduction 3](#_Toc127793313)

[What do I already know about the topic? 3](#_Toc127793314)

[Explain the problem under study 3](#_Toc127793315)

[What do I have to say critically about what is already known? 3](#_Toc127793316)

[Has anyone else done anything exactly the same? 3](#_Toc127793317)

[Has anyone else done anything that is related? 3](#_Toc127793318)

[Where does my work fit in with what has gone before? 4](#_Toc127793319)

[Why is my research worth doing in light of what has already been done? 4](#_Toc127793320)

[List the implication of answering the tentative research questions 4](#_Toc127793321)

[Data Introduction 5](#_Toc127793322)

[Dataset Description 5](#_Toc127793323)

[Record Removal & Value Imputation 5](#_Toc127793324)

[Fields To Be Dropped 5](#_Toc127793325)

[Fields To Be Retained 7](#_Toc127793326)

[GitHub 9](#_Toc127793327)

[Tentative Methodology 9](#_Toc127793328)

[References 9](#_Toc127793329)

# Topic Introduction

## What do I already know about the topic?

I am currently employed as a Business Intelligence Analyst by Innomar Strategies which is a healthcare company that focuses on providing access to specialty medications for a variety of disease states. I excel in providing a bridge between technical associates (Software Developers etc.), Subject Matter Experts (Health Care Professionals, Pharmaceutical Manufacturers) and those with little no exposure to the topic. Although I am familiar with healthcare terminology and the environment in general, I have no specific formal training in healthcare or the area of cardiac health.

## Explain the problem under study

My goal will be to predict based whether or not a participant in this survey has had a heart attack or any future person with the same set of data collected has/will have a heart attack. I will be focusing on independent variables within the study that I deem to be under the participant’s reasonable control. As a simple goal, I am looking to understand whether we can predict a heart attack with information provided that a user has control over and thus could change to potentially reduce the likelihood of a heart attack. I will not be focusing on information outside of a participants control and thus will remove these types of variables in the data cleansing process (e.g., is the survey respondent blind, has the survey respondent had previous traumatic experiences etc.).

## What do I have to say critically about what is already known?

There is a large disparity of knowledge on the topic globally (Ahmed, et al., n.d.). Studies have a large amount of Jargon that can be poorly understood by a layperson (e.g., simple example of Myocardial Infarction or MI meaning Heart Attack). I want to study the elements of a person’s habits/health that are commonly understandable and modifiable. For example, knowing a particular blood marker is associated with increased rate of heart attack may not help a layperson, but knowing a reduction in alcohol consumption may assist cardiac health may.

## Has anyone else done anything exactly the same?

While there are many studies completed on similar topics (including multivariate analysis to predict cardiac health) I have not found any examples of this exact study on this same data source.

## Has anyone else done anything that is related?

Although a large number of studies focus on post myocardial infarction outcomes, there are many studies and bodies of work that are related to my topic of prediction.

The topic can be approached from multiple angles (lifestyle choices vs genetics), multiple assumptions (healthy vs unhealthy individuals) and multiple different granularities (e.g., Assuming chronic inflammation may be associated with a higher level of MI, studies can focus entirely on what causes elevated inflammation levels), preventative medication in general or other potential variables that may otherwise not be considered due to their specificity, for example the study on noise exposure and risk of myocardial infarction incidence and mortality (Liu, Yan, Li Zou, & Fu, n.d.).

Of the studies reviewed, INTERHEART(YUSUF & RANGARAJAN, n.d.)was a highly similar observational study that I have identified. Using 12,152 MI cases and 14,820 control cases from across the globe, this study identified 9 easily measurable and modifiable risk factors could explain more than 90 percent of the risk of a heart attack. Smoking, consumption of fruits and vegetables, consumption of alcohol and physical activity were all associated with increasing the risk of Myocardial Infarction and similar variables will be included in my project.

## Where does my work fit in with what has gone before?

My goal is to analyse factors that are either within the control of the patient/user or is easily understandable by the patient/user. My diversity of information geographically and study design is much weaker compared to the INTERHEART study, however the breadth of information is larger from the Behavioural Risk Factor Surveillance System (BRFSS), and so it will be interesting to understand and observe how the outcomes of my project compares to the INTERHEART study.

## Why is my research worth doing in light of what has already been done?

According to the World Health Organization Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year(Cardiovascular diseases, n.d.). This makes this area of research incredibly important, and any contribution regardless of how minute to the topic potentially invaluable.

## List the implication of answering the tentative research questions

If I am able to successfully answer my tentative research question and predict the incidence of a heart attack based on lifestyle/modifiable variables within a survey participants response then the implication would be that for other humans, the risk of a heart attack can be reduced by modifying these behaviours.

# Data Introduction

## Dataset Description

### Record Removal & Value Imputation

In order to reduce the likelihood of introducing bias to the analysis the only records that will be removed entirely are those where DISPCODE is not equal to 1200 (indicating partial survey completion) and those where CVDINFR4 was not answered (since this is our predicted variable).

For individual missing elements within fields, these will be imputed based on mode, mean or median depending on the nature of the variable (e.g., categorical vs integer) as well as the distribution of the variable. There are variables where a blank response however should remain e.g. Does a participant drink alcohol? Where ‘No’ there would be no response for the number of drinks per day and would not be imputed with any value.

### Fields To Be Dropped

Because this dataset is derived from a general health survey and not specific to cardiac health or potential indicators of this a large number of fields can be removed. This will assist in focusing on the specific goal of this project, reduce initial computational complexity and also reduce the likelihood of overfitting to the training data.

|  |  |
| --- | --- |
| **Variable Name** | **Reason** |
| DISPCODE,\_STATE,FMONTH, IDATE, IMONTH, IDAY, IYEAR, SEQNO, \_PSU, CTELENM1, PVTRESD1, COLGHOUS, STATERE1, CELPHON1, LADULT1, NUMADULT, NUMMEN, NUMWOMEN, RESPSLCT, SAFETIME, CTELNUM1, CELLFON5, CADULT1, PVTRESD3, CCLGHOUS, CSTATE1, LANDLINE, HHADULT, RENTHOM1, FIREARM5, GUNLOAD, LOADULK2, QSTVER, QSTLANG, \_METSTAT, \_URBSTAT, MSCODE, \_STSTR, \_STRWT, \_RAWRAKE, \_WT2RAKE,\_CLLCPWT, \_DUALUSE, \_DUALCOR,\_LLCPWT2, \_LLCPWT | These are administrative type fields indicating items like the date of survey, landline vs cellphone, survey language geographic area within USA etc. or fields where a connection would be highly unlikely in my own opinion to be causal e.g. owning firearms |
| DIABAGE3, ARTHEXER, ARTHEDU, ARTHDIS2, DIFFWALK, DIFFDRES, DIFFALON, FLSHTMY3, IMFVPLA2, PNEUVAC4, HIVTST7, HIVTSTD3, PDIABTST, FEETCHK3, FEETCHK, EYEEXAM1, DIABEYE, DIABEDU, DOCTDIAB, HPVADVC4, HPVADSHT, TETANUS1, SHINGLE2, LCSCTSCN, HADMAM, HOWLONG, CERVSCRN, CRVCLCNC, CRVCLPAP, CRVCLHPV,PSATEST1,PSATIME1,PCPSARS2, PCSTALK, HADSIGM4 COLNSIGM, COLNTES1 SIGMTES1 LASTSIG4, COLNCNCR, VIRCOLO1, VCLNTES1, SMALSTOL STOLTEST STOOLDN1, BLDSTFIT, SDNATES1, CAREGIV1, CRGVREL4, CRGVLNG1, CRGVHRS1, CRGVPRB3, CRGVALZD, CRGVPER1, CRGVHOU1, CRGVEXPT, RCSGENDR, RCSRLTN2, CASTHDX2, CASTHNO2, STOPSMK2, HADHYST2, TRNSGNDR, SOFEMALE, SOMALE, ACEADNED, ACEADSAF, ACEHVSEX, ACETTHEM, ACETOUCH, ACESWEAR, ACEHURT1, ACEPUNCH, ACEDIVRC, ACEPRISN, ACEDRUGS, ACEDRINK, ACEDEPRS, CDDISCUS, CDSOCIAL, CDHELP, CDASSIST, CDHOUSE, CIMEMLOS, SHAREBP, WHEREBP, HOMRGCHK, HOMBPCHK, CSRVCTL2, CSRVPAIN, CSRVCLIN, CSRVDEIN, CSRVINSR, CSRVINST, CSRVRTRN, CSRVSUM, CSRVDOC1, CSRVTRT3, CNCRTYP1, CNCRAGE, CNCRDIFF, CVDCRHD4, CVDSTRK3, ASTHMA3, ASTHNOW, CHCSCNCR, CHCOCNCR, CHCCOPD3, ADDEPEV3, CHCKDNY2, DIABETE4, HAVARTH5, LMTJOIN3, JOINPAI2, DEAF, BLIND, DECIDE, PREDIAB1, INSULIN1, CHKHEMO3, TOLDCFS, HAVECFS, WORKCFS, PERSDOC3, FLUSHOT7, PNEUVAC4, MARITAL, PRIMINSR | These are other medical fields that are either:   * Correlated with other variables (which are not being removed). E.g., Someone with cancer will likely rate general health lower. * Directly linked to a different medical condition or screening for that condition e.g., physical recommendations for someone with Arthritis * Not something that can reasonably be modified by user e.g. To reduce your chance of MI don’t be blind |
| EMPLOY1 | Would otherwise be relevant fields, however utilizing another highly correlated field e.g. Are you employed compared to annual income |
| TOLDHEPC, TRETHEPC, PRIRHEPC, HAVEHEPC, HAVEHEPB, MEDSHEPB, WTCHSALT, DRADVISE, BIRTHSEX | Variables that have no variation and therefore do not assist me with predictive abilities. E.g. are either mostly blank or same value |
| \_CHISPNC,\_CRACE1, \_CPRACE1, CAGEG, \_RFHLTH, \_PHYS14D, \_MENT14D, \_HLTHPLN, \_HCVU652, \_TOTINDA, \_RFHYPE6, \_CHOLCH3, \_RFCHOL3, \_MICHD, \_LTASTH1, \_CASTHM1, \_ASTHMS1, \_DRDXAR3, \_DRDXAR3, \_LMTWRK3, \_PRACE1, \_MRACE1, \_HISPANC, \_RACE, \_RACEG21, \_RACEGR3, \_RACEPRV, \_SEX, \_AGEG5YR, \_AGE65YR, \_AGE80, \_AGE\_G, HTM4, \_BMI5, \_RFBMI5, \_CHLDCNT, \_EDUCAG, \_RFSMOK3, DRNKANY5, DROCDY3\_, \_RFBING5, \_DRNKWK1, \_RFDRHV7, \_FLSHOT7, \_PNEUMO3, \_AIDTST4, \_MISFRT1, \_MISVEG1, \_FRTRES1, \_VEGRES1 \_LMTACT3, HTIN4, \_MISVEG1, \_SMOKER3 | Calculated variables that I do not require either due to one of the above reasons, or due to utilizing the raw data instead of calculated |
| WEIGHT2, HEIGHT3, INCOME3, 'COLGSEX', 'LANDSEX', 'CELLSEX', 'SMOKE100', 'SMOKDAY2', FRUIT2, FRUITJU2, FVGREEN1, FRENCHF1, POTATOE1, VEGETAB2 | Variables that are useful, however I have opted for the calculated variable instead due to ease of use |
| FTJUDA2\_, FRUTDA2\_, GRENDA1\_, FRNCHDA\_, POTADA1\_, VEGEDA2\_, \_FRUTSU1, \_VEGESU1, \_FRTLT1A, \_VEGLT1A, \_FRT16A, \_VEG23A, \_FRUITE1, \_VEGETE1, USEMRJN3, DRNK3GE5 | After reviewing the data’s codebook [[1]](#footnote-1)this data is still considered by myself to be too ambiguous to use, despite my strong desire to do so due to previous study outcomes (noted on below table under ‘Diet’ |

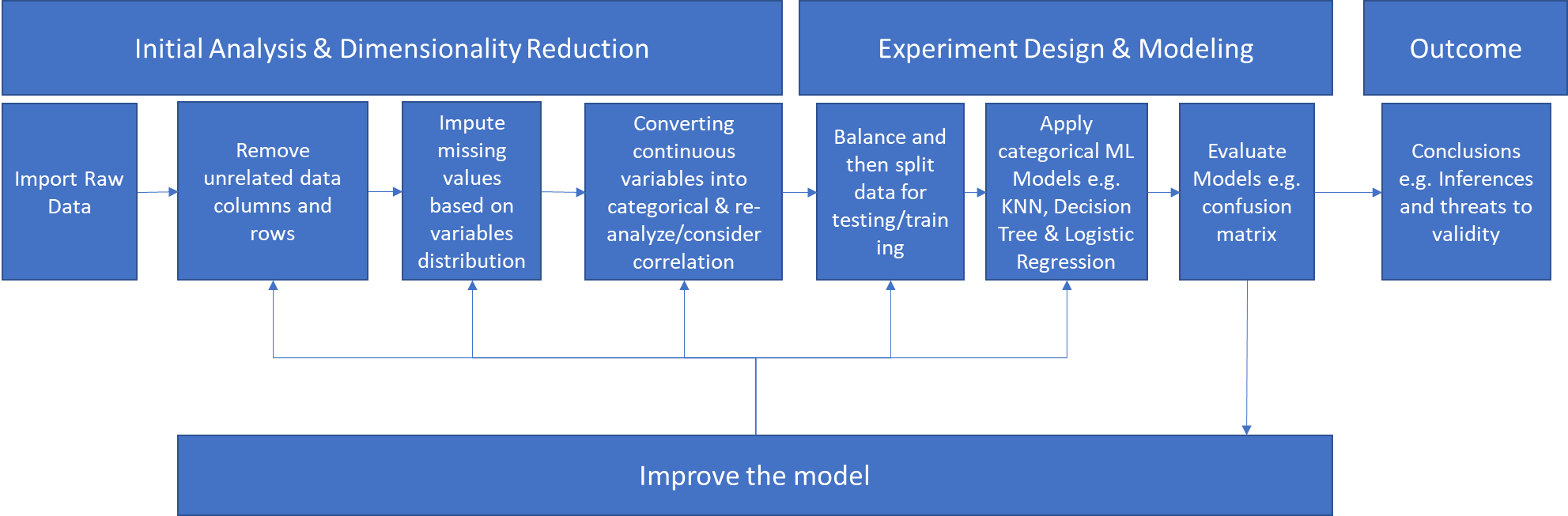
### Fields To Be Retained

|  |  |  |
| --- | --- | --- |
| **Group** | **Variable** | **Explanation/Reference** |
| Diet | *Note: While my research has shown food consumption is suggested to have a significant impact on cardiovascular health, the way the variables are set up in the study is too ambiguous for me to use, despite reviewing the full definition and way in which the question was asked. I am leaving here to show that this variable has been considered and studies reviewed.* | Food Consumption and its impact on cardiovascular disease: Importance of Solutions focused on the globalized food system (Anand, et al., n.d.) |
| Exercise | EXERANY2 | Walking more than 4 hours a week may reduce risk of hospitalization for cardiovascular disease events (PhD, PhD, PhD, MS, & Edward H. Wagner MD, n.d.) |
| Tobacco Use | USENOW3, ECIGNOW1, LCSFIRST, LCSLAST, LCSNUMCG | Studies indicate that smoking can increase the risk of developing heart failure (Long-Term Study Finds Cigarette Smoking Doubled Risk of Developing Heart Failure, n.d.) |
| Marijuana Use | MARIJAN1, RSNMRJN2 | While I am not expecting this to be significant, there are studies around short-term effects on risk of MI after Marijuana use (Mittleman, Lewis, Maclure, Sherwood, & Muller, n.d.) |
| Alcohol Use | ALCDAY5, AVEDRNK3, MAXDRNKS | Effect of Alcohol on Risk of Coronary Heart Disease and Stroke: Causality, Bias, or a Bit of Both? (Emberson & Bennett, n.d.) |
| Demographic | SEXVAR, MEDCOST1, EDUCA, \_IMPRACE, \_BMI5CAT, \_INCOMG1, | Demographical information will be useful to ensure a representative sample. Some demographic information such as BMI can be linked to increase incidence of MI based on INTERHEART study |
| General Health | GENHLTH, PHYSHLTH, MENTHLTH, POORHLTH, BPHIGH6, BPMEDS, CHOLCHK3, TOLDHI3, CHOLMED3 | While these are not directly modifiable, if these are associated with higher or lower incident of MI then common sense will allow these to be improved in most instances (E.g., research how to improve physical or mental health) |

## GitHub

<https://github.com/faulknbenj/Cind820>

## Methodology



## Analysis Limitations

**Bias in dataset:** A large limitation in the dataset is that I do not know when a heart attack has happened and therefore whether the lifestyle indicators that I am using to predict a heart attack are truly predictors of a heart attack, or are perhaps the result of a heart attack. For example – Did low activity increase chances of a heart attack, or has the individual reduced activity because of their previous heart attack. I believe this is a significant issue when interpreting my results and arguably invalidates any real-world application of the project.

**T-Test assumptions:** When utilizing the T-test as a method of evaluating the ‘best’ algorithm I have potentially breached the assumption that a large sample size is used. I have a Cross Validation running across 10 folds, so n=10, whereas a sample size of n>=30 would be preferential. However, due to performance constraints/concerns, I have not run the cross validation with this number of folds.

**Data Integrity:** I was unable to utilize any of the diet variables that I had hoped to include, this was a significant disappointment to me as previous studies (mentioned above) suggest diet has an impact on the likelihood of a heart attack. These data elements did not make sense in their interpretation based off their provided definition, and I therefore could not be sure that including these would improve the model, particularly when utilized on separate test data or real world application.

**Efficiency:** As noted below in section ‘What would I do differently next time’ I have noted that a large amount of time is spent by the computer looping through my data in order to categorize various elements into groups. Looking back I believe this may be unnecessary and potentially harmful to the algorithms ability to learn. I also have efficiency concerns in balancing the data, due to the comparatively rare indicator of a heart attack, this as a time consuming process to synthetically increase the number of heart attack respondents in the dataset.

**Ethical Considerations:** I do not believe there are any ethical considerations to be made regarding the utilization of this data set. This information has been voluntarily and anonymously provided, with the knowledge that the data set was a public data set. Although minimal, the only ethical consideration I would is in the practical use of the final product/algorithm. If it was to be used as a ‘risk calculator’ by individuals, could a false positive increase stress in a user and become a self-fulfilling prophecy to cause a heart attack

## Project Conclusion

The most accurate model, by a statistically significant amount, was the decision tree model with an accuracy of 62.6% on training data. Although throughout the project I have considered other measures of ‘success’ (recall, precision) via confusion matrix, I have used accuracy as my final ‘most important’ indicator of success as I do not feel as though false positives or false negatives have a disproportionate impact and therefore maximising accuracy on a balanced data set was my top priority.

When the decision tree algorithm is run using test data (data that it has never seen before) an accuracy of 61.9% is achieved. This has been shown to be a statistically significant reduction in accuracy (pvalue=0.0007) which may suggest some overfitting to the training dataset. However, this is still a relatively close accuracy outcome (.7% absolute difference).

## What would I do differently

Data grouping: I would refrain from grouping the data prior to training my algorithms and then potentially group after training has occurred and compare results to see if this has an impact. My concern with the layout of my current project is that I have reduces the algorithms ability to see the granularity of the data, and this may have had an impact on its predictions. My intention with this was to reduce the chance of overfitting, and to increase the readability of the decision tree output (however complexity is still much too high to have a ‘usable’ decision tree graphic to most users).

## Next Steps

I am uncertain whether I would consider a 62.6% accuracy to be a good outcome for this model. Although the algorithm is making decisions and is doing better than chance alone, I am not entirely satisfied with the final accuracy.

If I take this project further, more research is required from myself in order to understand how to best tune these algorithms and to understand any further classification algorithms that may be worth considering (e.g. Random Forest or Neural Networks). Although I have identified Decision Tree model as being more accurate than KNN and Logistic Regression, I do not feel as though I am at the stage where I would abandon these algorithms for the sole pursuit of improving the Decision Tree algorithm.

The study I have trained my algorithm on is an annual study, so it would be of particular interest to run this algorithm on a new version of the study to see how it performs on this ‘new’ data.

# References

Ahmed, A. A., AL-Shami, A. M., Jamshed, S., Zawiah, M., Elnaem, M. H., & Ibrahim, M. I. (n.d.). *Awareness of the Risk Factors for Heart Attack Among the General Public in Pahang, Malaysia: A Cross-Sectional Study*. Retrieved from https://www.tandfonline.com/doi/full/10.2147/RMHP.S281285

Anand, S. S., Hawkes, C., Souza, R. J., Mente, A., Dehghan, M., Nugent, R., . . . Martin, M. A. (n.d.). *Food Consumption and its impact on Cardiovascular Disease: Importance of Solutions focused on the globalized food system*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4597475

*Cardiovascular diseases*. (n.d.). Retrieved from World Health Organization: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab\_1

Emberson, J. R., & Bennett, D. A. (n.d.). *Effect of Alcohol on Risk of Coronary Heart Disease and Stroke: Causality, Bias, or a Bit of Both?* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1993990/

Liu, Y., Yan, S., Li Zou, J. W., & Fu, W. (n.d.). *Noise exposure and risk of myocardial infarction incidence and mortality: a dose–response meta-analysis*. Retrieved from https://link-springer-com.ezproxy.lib.torontomu.ca/article/10.1007/s11356-022-20377-w

*Long-Term Study Finds Cigarette Smoking Doubled Risk of Developing Heart Failure*. (n.d.). Retrieved from https://publichealth.jhu.edu/2022/long-term-study-finds-cigarette-smoking-doubled-risk-of-developing-heart-failure#:~:text=Long%2DTerm%20Study%20Finds%20Cigarette%20Smoking%20Doubled%20Risk%20of%20Developing%20Heart%20Failure,-Large%20study%20also&text=A%

Mittleman, M. A., Lewis, R. A., Maclure, M., Sherwood, J. B., & Muller, J. E. (n.d.). *Triggering Myocardial Infarction by Marijuana*. Retrieved from https://www.ahajournals.org/doi/10.1161/01.CIR.103.23.2805#:~:text=The%20risk%20of%20myocardial

PhD, A. Z., PhD, S. G., PhD, J. A., MS, L. C., & Edward H. Wagner MD, M. (n.d.). *Does Walking Decrease the Risk of Cardiovascular Disease Hospitalizations and Death in Older Adults?* Retrieved from https://agsjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.1532-5415.1996.tb02425.x

YUSUF, S., & RANGARAJAN, S. (n.d.). *INTERHEART - Global Risk Factors for Acute Myocardial Infarction*. Retrieved from https://www.phri.ca/research/interheart/

1. https://www.cdc.gov/brfss/annual\_data/2021/pdf/codebook21\_llcp-v2-508.pdf [↑](#footnote-ref-1)