Study  Protocol:

Development  and  internal  validation  of  a  NYHA  functional

class  prediction  algorithm  for  patients  with  (severe  to

moderate)  heart  failure

Jonathan-F.  Baril

November  11,  2017

1

**Abstract**

BACKGROUND:  Care  for  Heart  Failure  while  already  challenging  is  made  even  more  diﬃcult

as  there  is  no  reliable  objective  method  for  assessing  the  moment-to-moment  state  of  given

patients  HF,  never  mind  determining  if  it  is  likely  to  worsen.  Finding  an  objective,  or  at  least

consistently  replicable,  means  of  determining  a  patient’s  NYHA  class  would  provide  a  great

boon  to  both  heart  failure  care  and  research  as  it  would  cause  intra-  and  inter-physician  and

patient  assessments  of  heart  failure  class  to  more  consistent  making  communication  of  patient

heart  failure  states  in  research,  clinic  notes,  or  other  medical  documentation  more  transparent,

reliable  and  generalizable.

SCIENTIFIC  OBJECTIVES:

1.  Establish  a  robust  and  general  foundational  framework  for  developing  and  assessing  in-

telligent  NYHA  machine  classiﬁcation  algorithms  that  can  be  trained  to  mirror  grading

by  experienced  physicians  with  the  goal  of  making  NYHA  classiﬁcation  more  consistent

and  reliable  for  the  medical  research  and  clinical  community.

2.  To  perform  a  pilot  study  using  this  framework  to  determine  if  Hidden  Markov  Models

show  promise  for  NYHA  classiﬁcation.

STUDY  QUESTIONS:  Are  Hidden  Markov  Models  or  common  variants  such  as  Hidden  Semi-

Markov  Models  a  promising  method  of  performing  reliable,  objective  and  consistent  NYHA

Classiﬁcation?

HYPOTHESIS:  The  developed  Hidden  Markov  Model  based  classiﬁer  will  have  an  ’inter-rater

reliability’  within  the  range  of  or  greater  than  the  prevailing  norm  amongst  human  physicians

(50%-75%)  when  comparing  it’s  classiﬁcations  in  the  test  data  labelled  by  an  expert  physician

[[43](#link_bookmark_29)–[47](#link_bookmark_30)].

DESIGN  METHODOLOGY:  Sets,  composed  of  one  model  for  each  target  NYHA  class  (II  &

II),  of  multivariate  hidden  (pure-  or  semi-)  Markov  models.

SETTING:  Tertiary  care  clinic  specializing  in  the  management  of  heart  failure  (The  Heart

Function  Clinic  at  Toronto  General  Hospital,  a  part  of  the  University  Health  Network  (UHN)

in  Toronto,  Canada)  from  November  2017  to  April  2017.

PARTICIPANTS:  consecutive  sample  of  n  consenting  heart  failures  patients  participating  in

program  evaluation  and  quality  improvement  eﬀorts  for  Medly,  a  telemonitoring  platform  used

and  developed  for  the  test  site.  Patient  data  collected  using  digital  questionnaires,  clinic  notes

and  ﬁtness  tracker  (Fitbit  Charge  HR).

PRIMARY  OUTCOME:  NYHA  functional  classiﬁcation  as  assessed  by  patient’s  attending

cardiologist  at  clinic  visits  during  study  period.

RESULTS:  TBD

CONCLUSIONS:  TBD

[Classiﬁcation;  Development;  Internal  Validation]

2

**Contents**

[**1 Introduction**](#Introduction)

[**4**](#Introduction)

[1.1   Background](#Background___Objectives)  [& Objectives  .](#Background___Objectives)

[4](#Background___Objectives)

[**2 Methods**](#Methods)

[**6**](#Methods)

[2.1   Source](#Source_of_Data/Study_Design)  [of  Data/Study](#Source_of_Data/Study_Design)  [Design](#Source_of_Data/Study_Design)   [.  .](#Source_of_Data/Study_Design)

[6](#Source_of_Data/Study_Design)

[2.2   Participants .](#Participants)

[6](#Participants)

[2.3   Outcome](#Outcome)  [.  .  .](#Outcome)

[7](#Outcome)

[2.3.1     Risk](#Risk_Groups)  [Groups](#Risk_Groups)

.  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .

[7](#Risk_Groups)

[2.3.2     Label  Assignment](#Label_Assignment)

.  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .

[8](#Label_Assignment)

[2.4   Predictors](#Predictors)  [.  .  .  .10](#Predictors)

[2.5   Sample](#Sample_Size)  [Size .12](#Sample_Size)

[2.6   Missing](#Missing_Data)  [Data](#Missing_Data)   [.  .13](#Missing_Data)

[2.7   Statistical](#Statistical_Analysis_Methods)  [Analysis](#Statistical_Analysis_Methods)  [Methods](#Statistical_Analysis_Methods) [. 13](#Statistical_Analysis_Methods)

[2.7.1     Normalization](#Normalization) [.  .  .  .  .13](#Normalization)

[2.7.2     Classiﬁer](#Classifier_(Model)_Development)  [(Model)](#Classifier_(Model)_Development)  [Development .14](#Classifier_(Model)_Development)

[2.7.3     Classiﬁer](#Classifier_(Model)_Performance_Assessment)  [(Model)](#Classifier_(Model)_Performance_Assessment)  [Performance](#Classifier_(Model)_Performance_Assessment)  [Assessment .29](#Classifier_(Model)_Performance_Assessment)

[2.8   Development vs.](#Development_vs._Validation)  [Validation](#Development_vs._Validation)   [.  .  .  41](#Development_vs._Validation)

[**3 Results**](#Results)

[**41**](#Results)

[3.1   Participants .41](#Participants)

[3.2   Model](#Model_Development)  [Development](#Model_Development)

.  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .    [41](#Model_Development)

[3.3   Model](#Model_Specification)  [Speciﬁcation](#Model_Specification)  [.  .  .  .41](#Model_Specification)

[3.4   Model](#Model_Performance)  [Performance](#Model_Performance)  [.  .  . 41](#Model_Performance)

[3.5   Model-updating](#Model-updating)  [.  .  41](#Model-updating)

[**4 Discussion**](#Discussion)

[**41**](#Discussion)

[4.1   Limitations](#Limitations)   [.  .  .  . 41](#Limitations)

[4.2   Interpretation](#Interpretation)  [.  .  .  .  41](#Interpretation)

[4.3   Implications](#Implications)  [.  .  .  .  41](#Implications)

[**5 Other**](#Other_Information)[**Information**](#Other_Information)

[**41**](#Other_Information)

[5.1   Supplementary](#Supplementary_Information)  [Information](#Supplementary_Information)   [.  .  .  . 41](#Supplementary_Information)

[5.2   Funding](#Funding)   [.  .  41](#Funding)

3

**1     Introduction**

**1.1     Background  &  Objectives**

Heart failure, the complex chronic terminal phase of all cardiovascular disease, is slowly becoming

a worldwide silent pandemic [[1](#link_bookmark_89)].  The symptoms of heart failure are complex and diﬃcult to manage

for  both  patients  and  their  physicians  [[2](#link_bookmark_90)–[4](#link_bookmark_91)].  Care  is  made  even  more  diﬃcult  because  there  is  no

reliable  objective  method  for  assessing  the  moment-to-moment  state  of  given  patients  HF,  never

mind determining if it is likely to worsen [[5](#link_bookmark_92)–[7](#link_bookmark_93)].  The current clinical gold standard for communicating

a patients state of Heart Failure is the New York Heart Association (NYHA) functional classiﬁcation

[[8](#link_bookmark_94), [9](#link_bookmark_95)].  This  system  grades  a  patient’s  degree  of  heart  failure  based  on  the  physicians  interpretation

of  patient  reported  symptoms  (mainly  with  respect  to  their  degree  of  exercise/activity  intolerance).

Despite  these  limitations  clinical  evidence  and  medical  research  have  established  many  important

relationships  between  a  patient’s  NYHA  class  and  their  prognostic  outcomes  [[7](#link_bookmark_93), [10](#link_bookmark_97)].

Finding  an  objective  means  of  determining  a  patient’s  NYHA  class  would  provide  a  great  boon

to  both  heart  failure  care  and  research  as  it  would  cause  intra-  and  inter-physician  and  patient

assessments  of  heart  failure  class  to  more  consistent  [[7](#link_bookmark_93), [11](#link_bookmark_99), [12](#link_bookmark_100)].  Doing  so  would  make  communica-

tion  of  patient  heart  failure  states  in  research,  clinic  notes,  or  other  medical  documentation  more

transparent  and  reliable.

Subjectivity  in  the  current  NYHA  classiﬁcation  is  introduced  through  two  primary  sources:

patients  and  clinicians.  Clinician,  who  readily  identify  these  as  issues,  have  already  made  steps  to

try  and  determine  ways  to  eliminate  this  subjectivity  [[7](#link_bookmark_93)].   One  of  the  common  clinical  tests  used

is  the  6  minute  walk  test  (6MWT),  where  a  patient  is  asked  to  walk  as  far  as  they  can  (being

permitted  to  rest  as  needed)  over  a  hard  ﬂat  surface  over  the  period  of  6  minutes;  the  total  distance

walked  is  then  used  as  an  indicator  of  the  aerobic  capacity  of  the  individual  and  by  inference  their

degree  of  heart  failure  [[13](#link_bookmark_102)].   The  development  of  the  6MWT  typiﬁes  the  general  approach  being

taken  to  improve  this  particular  area  of  clinical  practice.  Much  research  revolves  around  trying  to

identify  or  create  tests  that  measure  physical  ﬁtness,  maximum  exercise  capacity  or  some  proxy

thereof  [[14](#link_bookmark_103)–[21](#link_bookmark_104)].  These  tests  are  then  used  to  infer  a  patient’s  symptom  response  to  exercise  in  their

everyday  life  in  order  to  compare  it  to  the  NYHA  classiﬁcation  of  that  same  patient.  Typically  a

high  level  of  physical  ﬁtness  is  assumed  to  imply  less  exercise  intolerance  in  everyday  life,  which

then implies a lower NYHA classiﬁcation [[14](#link_bookmark_103),[18](#link_bookmark_106),[21](#link_bookmark_104)].  These tests are,  almost without exception,  run

in  a  controlled  clinical  environment  and  are  supervised  by  trained  staﬀ.  As  a  result  they  generally

do  not  measure  the  patients  real  world  response  to  physical  exertion,  in  contrast  to  the  aim  of  the

NYHA  functional  classiﬁcation.

In  general  these  tests  revolve  around  measuring  a  patient’s  exertion  over  a  period  of  time  [[13](#link_bookmark_102),

[14](#link_bookmark_103), [16](#link_bookmark_110)–[18](#link_bookmark_106), [20](#link_bookmark_112), [21](#link_bookmark_104)].   Exertion  is  usually  calculated  by  raw  distance  traveled  (being  generally  more

convenient  to  measure)  [[13](#link_bookmark_102), [14](#link_bookmark_103), [16](#link_bookmark_110), [18](#link_bookmark_106)],  patient  step  count  (which  can  be  linked  to  distance  if  the

patient’s  stride  length  is  known)  [[20](#link_bookmark_112),[22](#link_bookmark_119)–[28](#link_bookmark_120)],  movement  recorded  by  raw  accelerometer  data  [[21](#link_bookmark_104),[29](#link_bookmark_122)–

[31](#link_bookmark_123)],  or  even  diﬃculty  (e.g.  surface  incline,  resistance  band  strength)  [[22](#link_bookmark_119),[27](#link_bookmark_125)]  or  energy  consumption

(e.g.  Metabolic  Equivalents:  METS)  [[8](#link_bookmark_94), [14](#link_bookmark_103), [19](#link_bookmark_128)]  of  exercise  being  attempted.

While  these  tests  have  shown  that  measures  of  exertion  over  time  (whether  distance,  step  count

or  otherwise)  are  correlated  to  the  NYHA  functional  classiﬁcation  of  patients  there  often  remains

a  notable  gap  in  the  explanatory  power  of  these  measures.  For  example  Demers  et  al.  found  that

for  the  768  patients  in  their  multi-centre  study  the  “baseline  6MWT  distance  was  ...   moderately

inversely  correlated  to  the  New  York  Heart  Association  functional  classiﬁcation  (NYHA-FC)  (r  =  -

0.43, P =.001)” [[32](#link_bookmark_129)].  In other words, distance travelled seemed to only explain approximately 18.5%

4

of  the  variance  in  the  data  (r    =  0.1849).   This  is  probably  unsurprising  since  NYHA  functional

2

class  is  not  predominantly  attempting  to  ascertain  maximal  exercise  capacity  but  rather  the  degree

of  abnormally  symptomatic  response  to  exercise.  Therefore  tests,  measures,  or  metrics  which  can

reliably  mirror  NYHA  functional  class  will  likely  need  to  measure  not  just  exertion  but  patient

response  to  that  exertion  -  beyond  the  simply  binary  response  of  ’not  being  able  to  continue  the

exertion  demanded’  (the  case  for  all  the  previously  mentioned  tests).  In  fact,  it  is  our  hypothesis

that  when  attempting  to  relate  the  results  of  a  a  test  back  to  NYHA  functional  classiﬁcation,

tests  that  account  for  patient  response  to  exertion  will  have  superior  predictive  capacity  (whether

measured  by  correlation,  classiﬁcation  accuracy,  or  other  metric  as  appropriate  for  comparing  the

tests  in  question)  compared  to  those  tests  that  simply  measure  patient  exertion.

Heart  rate,  which  has  prognostic  value  by  itself,  is  an  obvious  and  simple  metric  that  can  be

used  to  assess  patient/cardiac  response  to  exercise  [[15](#link_bookmark_130), [33](#link_bookmark_131)–[36](#link_bookmark_132)].   As  an  example  then,  a  test  that

simultaneously  measures  heart  rate  along  with  an  exertion  measure  such  as  distance  traveled  or

step count would be much better positioned to provide an assessment of a patient’s speciﬁc response

to  exertion.

It  turns  out  that  modern  commercially  available  activity/ﬁtness  trackers  can  fairly  reliably

track  both  minute-by-minute  step  count  (the  exertion  measure)  as  well  as  heart  rate  (the  exertion

response).   These  devices  are  a  promising  means  of  providing  a  human  memory-independent  and

precise  (even  if  not  a  100%  accurate)  picture  of  a  patient’s  response  to  activity  [[19](#link_bookmark_128), [24](#link_bookmark_134)–[26](#link_bookmark_135), [37](#link_bookmark_136)–

[41](#link_bookmark_137)].   “Replacing”  patient  memory  with  activity  trackers  would  eliminate  a  signiﬁcant  source  of

subjectivity  and  potential  error  in  trying  to  determine  a  patient’s  level  of  activity  intolerance.

An  intelligent  classiﬁcation  algorithm  could  then  be  trained  to  mimic  expert  grading  by  an

experienced  “model”  physician  (and  to  compensate  for  inherent  biases  and  inaccuracies  in  the

recording sensors).  By imbuing artiﬁcial intelligence into such an algorithm so that it could translate

relevant  data  into  the  desired  clinical  outcome  (NYHA  classiﬁcation)  or  a  suﬃciently  equivalent

outcome  (an  ’NYH-AI’  or  ’NYHAI’  classiﬁcation  if  you  will)  we  could  provide  a  way  for  to  assess  a

patient’s  functional  classiﬁcation  in  an  objective,  universally-consistent  manner  that  still  leverages

the  advantages  and  beneﬁts  of  the  existing  ’traditional’  NYHA  classiﬁcation  method.   In  other

words  a  continuously  worn  activity  tracker  combined  with  an  intelligent  classiﬁcation  algorithm

executed  by  a  computer  could  remove  the  subjectivity  introduced  by  both  clinical  and  patients

providing  a  more  consistent  and  purely  objective  means  of  measuring  patient  response  to  exertion

and  assessing  NYHA  class.

The  general  the  objectives  of  this  study  are  as  follows:

1.  The  primary  goal  is  to  establish  a  robust  and  general  foundational  framework  for  use  by

researchers,  data  scientists  and  engineers  to  develop  and  assess  intelligent  NYHA  machine

classiﬁcation  algorithms  that  can  be  trained  to  mirror  grading  by  experienced  physicians  with

the  goal  of  making  NYHA  classiﬁcation  more  consistent  and  reliable  for  the  medical  research

and  clinical  community.

2.  A  secondary  goal  is  to  initiate  the  start  the  collection  of  continuously  monitored  activity  and

heart  data  from  heart  failure  patient  at  the  test  site  for  use  in  this  and  future  studies  and

algorithm  development.

3.  The  tertiary  aim  is  to  perform  an  pilot  study  using  data  collected  during  an  initial  brief

data  collection  period  as  well  as  the  foundational  framework  developed,  to  explore  if  Hidden

Markov  Models  show  promise  as  a  means  of  developing  an  intelligent  NYHA  classiﬁcation

algorithm.

5

**2     Methods**

**2.1     Source  of  Data/Study  Design**

The  pilot  study  data  to  be  used  for  development  of  the  Hidden  Markov  Model  based  classiﬁer

will  be  sourced  between  December  2017  and  April  2017  from  an  open  (prospective)  cohort  of

adult  outpatients  at  a  tertiary  care  clinic  specializing  in  the  management  of  advanced  heart  failure

(The  Heart  Function  Clinic  at  Toronto  General  Hospital,  a  part  of  the  University  Health  Network

(UHN),  in  Toronto,  Canada).   Since  patients  will  be  enrolled  on  a  rolling  basis  throughout  this

period,  patients  who  have  been  monitored  for  less  than  1  month  at  the  end  of  the  study  period  will

be  excluded  from  this  initial  pilot  development  data  set.

The  same  dataset  used  for  algorithm  development  will  be  used  for  internal  validation  of  the

algorithm  since  a  suitable  external  validation  dataset  is  not  currently  available.

**2.2     Participants**

The  cohort  will  consist  of  all  consecutive  patients  prescribed,  enrolled  and  using  (i.e.   being

monitored  by)  the  Medly  telemonitoring  program  (Medly)  consenting  to  be  included  in  program

evaluation  and  quality  improvement  eﬀorts  for  Medly.

A  patient  will  be  considered  for  inclusion  into  Medly  only  if  they  are:

1.  a  consenting  adult  (18+  years  of  age),

2.  diagnosed  with  heart  failure,

3.  followed  by  a  licensed  cardiologist  at  the  UHN  Heart  Function  Clinic  who  in  turn  bears  the

primary  responsibility  for  the  management  and  care  of  that  patients  heart  failure  diagnosis

(the  responsible  or  attending  cardiologist),

4.  suﬃciently  capable  of  speaking  and  reading  English,  or  having  an  informal  caregiver  (spouse,

parent,  etc.)  capable  of  the  same  so  as  to  both:

(a)  undergo  the  process  of  and  provision  of  informed  consent  for  participation  in  the  Medly

program

(b)  understand  and  follow  the  text  prompts  provided  by  the  Medly  patient-side  application

5.  capable  of  complying  with  the  use  of  Medly  (e.g.   capable  of  truthfully  answering  symptom

questions,  capable  of  safely  and  correctly  using  the  peripherals  such  as  the  weight  scale,

activity  tracker  and  blood  pressure  cuﬀ)

While there are no explicit exclusion criteria for the study,  we note that the decision to prescribe

or  exclude  a  patient  from  the  Medly  program  is  ultimately  up  to  the  professional  judgement  of  the

attending  cardiologist.   As  of  the  time  of  writing  a  total  of  6  attending  cardiologists  use  Medly

as  part  of  patient  care  although  one  of  the  cardiologists  (the  medical  director  of  the  clinic)  is

disproportionately  responsible  for  a  majority  of  the  patients  monitored.

During  the  study  period  patients  will  receive  no  special  treatment  with  regards  to  the  medical

interventions used and prescribed to them.  Their cardiologist are expected to continue treat patients

according  to  the  established  standard-of-care  at  the  Heart  Function  Clinic.  It  should  be  noted  that

the  standard-of-care  is  expected  to  continue  to  evolve  during  the  study  period  to  reﬂect  current

6

clinical  best  practice.  It  is  expected  that  the  Medly  platform  will  continue  to  be  updated  to  reﬂect

this  evolving  standard-of-care  and  will  receive  regular  features  updates,  upgrades  and  bug-ﬁxes  as

released  by  the  Medly  development  team  from  time  to  time.

It is also worth noting that as part of regular heart failure care patients may receive medications

(such as beta-blockers) which are known to aﬀect heart rate.  However, since this data is not available

in  a  conveniently  and  easily  accessible  format  it  will  not  be  included  in  the  dataset  at  the  present

time  due  to  time  constraints  although  we  do  discuss  how  it  might  be  included  for  the  sake  of

completeness.

**2.3     Outcome**

The  outcomes  of  interest  are  the  New  York  Heart  Association  (NYHA)  functional  class  for

each  patient.  While  competing  methods  of  reporting  the  severity  of  heart  failure  experienced  by  a

patient  exist,  the  NYHA  classiﬁcation  is  arguably  the  most  commonly  used  system  for  functional

classiﬁcation  of  heart  failure  and  many  well  established  relationships  have  been  identiﬁed  between

a  patient’s  NYHA  class  and  prognostic  outcomes  for  the  condition.

**2.3.1     Risk  Groups**

Patients  in  the  study  are  classiﬁed  into  one  of  four  risk  groups.   NYHA  class  I,  II  ,  III  &  IV.

As  a  specialized  tertiary  care  centre  the  Heart  Function  Clinic  rarely  sees  NYHA  class  I  patients

as  they  are  often  asymptomatic  with  regards  to  their  heart  failure,  or  at  least  rarely  require  the

specialized  level  of  care  oﬀered  by  the  clinic.   NYHA  class  IV  patients,  while  seen  at  the  clinic,

often  derive  little  beneﬁt  with  regards  to  physical  activity  monitoring  since  these  patients  are

known  to  be  often  bedridden  or  severely  activity  intolerant.   Classes  II  and  III  however  represent

a  classiﬁcation  challenge  since  the  dividing  line  between  mild  and  marked  activity  intolerance  can

sometimes  be  unclear  especially  given  the  inherent  subjective  nature  of  the  NYHA  classiﬁcation;

training a machine to classify between class II & II patients is therefore likely to produce the greatest

beneﬁt  to  clinical  and  medical  research  so  we  focus  on  these  two  risk  groups  in  this  study.

NYHA  class  is  typically  assessed  for  every  patient  with  known  cardiac  disease  (usually  ﬁrst

objectively  veriﬁed  through  the  use  of  some  sort  of  medical  imaging  modality).  It  is  then  reassessed

at  every  clinic  visit  by  the  physician  responsible  for  patient’s  care.   At  minimum,  the  physician

will  pose  questions  to  attempt  to  elucidate  the  patients’  degree  of  exercise  intolerance,  for  example:

“How  far  can  you  walk  before  becoming  short  of  breath?”,  or  “How  many  ﬂights  of  stairs  can  you

climb before needing to stop?”.  The physician will then select an NYHA class based on their clinical

experience,  professional  judgement  according  to  the  NYHA  class  deﬁnitions.  These  deﬁnitions  are

copied  below  for  the  reader’s  convenience  [[42](#link_bookmark_138)]:

I:  “Patients  with  cardiac  disease  but  without  resulting  limitation  of  physical  activity.  Ordinary

physical  activity  does  not  cause  undue  fatigue,  palpitation,  dyspnea,  or  anginal  pain.”

II:  “...   slight  limitation  of  physical  activity.   They  are  comfortable  at  rest.   Ordinary  physical

activity  results  in  fatigue,  palpitation,  dyspnea,  or  anginal  pain.”

III:  “...  marked  limitation  of  physical  activity.  They  are  comfortable  at  rest.  Less  than  ordinary

activity  causes  fatigue,  palpitation,  dyspnea,  or  anginal  pain.”

7

IV:  “Patients  with  cardiac  disease  resulting  in  inability  to  carry  on  any  physical  activity  without

discomfort.  Symptoms  of  heart  failure  or  the  anginal  syndrome  may  be  present  even  at  rest.

If  any  physical  activity  is  undertaken,  discomfort  is  increased.”

**2.3.2     Label  Assignment**

At  the  study  site  (the  Heart  Function  Clinic)  NYHA  class  is  assessed  for  each  patient  as  part

of  each  visit  at  the  Heart  Function  Clinic.   The  assessment  is  performed  by  the  physician  seeing

the  patient  (resident,  fellow  or  otherwise)  and  veriﬁed  by  the  attending  physician.   Within  the

context  of  this  study  it  means  that  outcome  measures  are  recorded  at  the  date  of  on-boarding  and

then  only  recorded  sporadically  during  the  study  period  as  patients  return  for  clinic  visits.   The

interval  between  regularly  scheduled  clinic  visits  varies  between  1-2  weeks  to  3-6  months.  Patients

with  more  severe  or  less  stable  conditions  will  visit  more  frequently.   As  well,  patient’s  who  are

undergoing  a  change  in  their  heart  failure  medication  (e.g.   up-  or  down-  titrating)  will  also  visit

more  frequency,  usually  a  few  times  within  a  period  of  a  few  weeks  as  the  medication  is  adjusted.

Since  patients  were  on-boarded  on  a  rolling  basis  over  a  period  of  5  months  there  is  no  guarantee

that  a  given  patient  (especially  the  more  stable  patients)  will  return  to  physically  visit  the  clinic

within  the  study  period  and  be  reassessed  for  any  potential  change  in  their  NYHA  class.

The  sporadicity  and  instability  of  the  outcome  measure  presents  a  signiﬁcant  challenge  to  this

study  for  several  reasons.  There  is  however  a  further  challenge  with  regards  to  NYHA  label  assign-

ment:  physicians  are  notoriously  unreliable  and  inconsistent  at  assessing  NYHA  class  in  the  ﬁrst

place  [[43](#link_bookmark_29)–[47](#link_bookmark_30)].  Caroll  et  al.  report  (bibliographic  reference  numbers  updated  to  reﬂect  ours):

[One  study]  used  two  physicians  to  estimate  NYHA  functional  class  in  75  patients  on

the  same  day  without  chronic  heart  failure,  reporting  an  interrater  reliability  of  56%

(weighted  kappa  =  0.41)  [[44](#link_bookmark_141)].   In  a  second  study,  two  cardiologists  assessed  the  same

50  chronic  heart  failure  patients  on  the  same  day  in  random  order,  observing  54%

agreement  in  NYHA  classes  [[46](#link_bookmark_142)].    In  a  third  study,  two  physicians  assigned  NYHA

class  to  56  patients  with  stable  angina  within  the  same  hour,  resulting  in  the  highest

reported  agreement  of  75%  [[45](#link_bookmark_143)].  Among  these  studies,  disagreement  by  more  than  one

functional  class  was  low  and,  for  the  most  part,  was  concentrated  on  determining  the

discrete  diﬀerences  between  Classes  II  and  III.  Taken  together,  the  reliability  of  the

NYHA  system  is  limited  in  the  few  trials  that  have  measured  it  directly.  [[43](#link_bookmark_29)]

Kubo  et  al.  came  up  with  a  standardized  NYHA  questionnaire  for  use  at  multi-centre  studies

based  on  the  published  deﬁnitions  by  the  American  Heart  Association  to  try  and  improve  intra-

rater  agreement  and  reliability  [[47](#link_bookmark_30)].   Even  using  this  questionnaire  they  reported  a  concordance

of  only  about  60%  between  the  3  independent  physician  graders  and  a  concordance  of  about  75%

for  the  (independent)  reviewers  [[47](#link_bookmark_30)].   Concordance  for  “repeat  grading  of  30  randomly  selected

questionnaires” was admittedly much higher at 90% but we doubt this last result is highly extensible

to  typical  clinical  practice  [[47](#link_bookmark_30)].

In  other  words,  in  half  of  the  studies  outlined  above  inter-physician  agreement  on  NYHA  class

fared  only  marginally  better  than  random  and  even  in  those  studies  where  physician  agreement

was  much  better  than  random  there  was  still  often  disagreement  in  essentially  1/4  of  the  cases.

Training  an  algorithm  on  data  that  is  suspect  to  this  degree  is  essentially  already  doomed  to  fail  if

the  objective  is  to  be  able  to  unequivocally  determine  the  true  underlying  NYHA  class  of  a  patient

-  the  fact  that  patient  class  can  change  between  clinic  visits  without  being  perceived  or  detected  by

8

the  rating  physician  (and  therefore  updated  for  our  training  set)  only  serves  to  doom  such  a  goal

even  more.

So  why  bother?  Well  it  is  an  unwritten  truth  in  engineering  practice  that  while  bad  standards

are  bad,  they  are  at  least  consistently  so  which  does  make  them  much  easier  to  ﬁx.   Everyone

can  agree  that  a  bad  standard  is  bad,  but  it’s  universal  use  and  acceptance  does  at  least  provide

some  value.  So,  the  initial  value  an  automated  NYHA  classiﬁcation  algorithm  will  not  necessarily

be  in  it’s  ability  to  unequivocally  determine  the  true  underlying  NYHA  class  of  a  patient  but

rather  that  it  rates  patients  in  a  universally  consistent  manner.  For  example  consider  the  Montreal

Cognitive  Assessment  (MoCA)  and  Mini  Mental  State  Examination  (MMSE),  both  screening  tests

used  for  detection  of  mild  cognitive  impairment  (a  suspected  precursor  to  Alzheimer’s  disease)  [[48](#link_bookmark_148)].

Both  MoCA  and  MMSE  test  results  have  their  limitations  with  regards  to  detection  of  degrees  of

cognitive  impairment  with  varying  sensitivity  between  the  tests  leading  some  researchers  to  try  and

develop  alternative  but  both  the  MoCA  and  MMSE  test  are  still  often  used  to  by  researchers  and

clinicians  to  communicate  the  degree  of  a  patient’s  impairment  because,  importantly,  they  provide

a  consistent  and  reliable  even  if  sometime  imprecise  way  of  assessing  patient  state  [[48](#link_bookmark_148)–[50](#link_bookmark_150)].

Along  this  vein,  by  creating  a  system  that  adequately  mimics  the  classiﬁcation  decisions  of  an

accepted  authoritative  source  (i.e  our  expert  clinician)  we  can  essentially  create  such  a  consistent

test  for  heart  failure  classiﬁcation  while  maintaining  the  existing  utility  of  NYHA  class.  Disclosing

this algorithm would make it possible for clinicians and researchers to use the algorithm as a ’virtual’

independent  rater  to  generate  a  common  and  completely  consistent  way  of  communicating  heart

failure  state  regardless  of  major  diﬀerences  in  clinician  training,  experience,  patient  context,  etc.

It  is  also  for  this  reason  that  the  data  collected  from  patients  for  use  in  this  study  is  limited  to

computer  harvested  data  (such  as  that  from  activity  trackers)  or  to  independently  veriﬁable  data

that  is  measured  according  to  a  universally  (or  near-universal)  accepted  standard  (such  as  patient

age, or prescribed beta-blocker dosage) - to ensure the algorithm results are as universally applicable

as  possible.

In  this  way,  the  fact  that  patient  class  may  change  between  clinic  visits  is  of  secondary  concern

since  it  can  be  rectiﬁed  later  as  it  speaks  more  to  the  ability  of  the  algorithm  to  predict  the

underlying  NYHA  class  rather  than  mimic  the  rating  source.    However,  in  the  interest  of  not

having completely unreliable source data, since we cannot guarantee NYHA functional class stability

between  patient  visits  for  a  given  patient,  we  propose  using  the  Medly  symptoms  questionnaire  as

a  proxy  outcome  and  as  a  means  of  detecting  suspected  drops/aggravations  in  NYHA  class.   As

part  of  Medly,  patients  are  prompted  every  morning  by  the  telemonitoring  system  to  answer  the

following  series  of  questions  to  ascertain  their  current  condition.   All  of  these  questions  have  a

clinical  relevance  to  a  patients  current  heart  failure  condition  and  so  have  the  potential  to  alert

us  of  a  potential  change  (i.e.  drop)  in  a  patient’s  class.  Note  that  questions  [7](#link_bookmark_151)-[11  on](#link_bookmark_152)

patient  has  answered  yes  to  any  of  question  previous  questions  ([1](#link_bookmark_153)-[6](#link_bookmark_154)).

1.  Have  you  fainted?

2.  Has  your  ICD  (Implantable  Cardiverter  Deﬁbrillator)  gone  oﬀ?

3.  Has  your  breathing  as  night  worsened?

4.  Do  you  have  more  chest  pain  than  usual?

5.  Are  you  more  tired  than  usual?

6.  Are  you  more  short  of  breath  than  usual?

9

7.  Are  your  ankles  swollen?

8.  Do  you  feel  that  your  heart  is  beating  unusually?

9.  Do  you  feel  lightheaded?

10.  Did  you  have  to  stop  any  of  your  usual  daily  activities  because  of  your  health?

11.  Enter  the  number  of  pillows  you  used  to  sleep.

While all of the questions target some speciﬁc symptom or set of symptoms experienced by heart

failure  patients,  questions  [1](#link_bookmark_153),  [2](#link_bookmark_156),  [4](#link_bookmark_157)-[6](#link_bookmark_154),  &  [8](#link_bookmark_159)-[10  sp](#link_bookmark_160)

an  increase  in  a  patients  exercise/activity  intolerance.  Since  a  patient’s  level  of  activity  intolerance

is  the  primary  determinant  of  their  NYHA  classiﬁcation,  an  increase  in  the  frequency  of  aﬃrmative

responses to these questions over a sustained period,  say 2 weeks,  would be a strong indicator that a

patient’s  NYHA  class  may  have  changed.  Baseline  response  frequency  can  be  determined  from  the

2  week  period  immediately  following  the  ﬁrst  clinic  visit  moving  to  the  2  week  period  immediately

preceding the last clinic visit for all following visits.  Absent of a conﬁrmation of NYHA class change

at  the  next  clinic  visit  a  statistically  signiﬁcant  change,  at  the  p  =  0.1  level  of  signiﬁcance,  from  the

baseline  frequency  of  response  would  be  a  reasonable  indicator  that  a  patient’s  NYHA  class  may

have  experienced  a  level  of  instability  of  the  2  week  period  worth  following  up  on.   To  follow-up

and  verify  a  possible  NYHA  class  change  patients  would  be  requested  to  complete  the  standardized

questionnaire  developed  by  Kubo  et  al.  [[47](#link_bookmark_30)].  The  completed  questionnaire  would  then  be  provided

to  the  clinician  responsible  for  that  patient  so  they  can  perform  a  reassessment  of  the  patients

NYHA  class.

The  patient  data  streams  will  be  segmented  into  two-week  chunks  with  each  chunk  representing

one  time-unit  of  analysis  so  that  the  algorithm  will  be  made  to  attempt  to  independently  classify

(and  trained  on)  each  of  these  two-week  chunks  of  data.

**2.4     Predictors**

The  predictors  used  to  generate  the  prediction  outcome  are  as  follows:

1.  Heart  Rate  [beats  per  minute]  -  recorded  using  commercially  available  activity-tracker  (Fitbit

Charge  HR,  Charge  2  or  Alta  HR)  continuously  throughout  the  day.

Fitbit  activity-trackers  measure  heart  rate  using  a  technique  known  as  photoplethysmograph

(PPG) [[22](#link_bookmark_119),[51](#link_bookmark_163)].  The devices are equipped with signal emitter(s) - light emitting diodes (LEDs),

generally  of  diﬀerent  frequencies  -  and  matched  signal  detector(s)  -  light  detecting  optical

sensors  matched  to  the  frequency  of  the  LEDs.  The  signal  emitter  and  detectors  are  placed

in  close  proximity  to  a  patient’s  skin  so  that  the  light  can  be  directed  to  shine  on  their  skin.

In  this  way  the  light  emitted  by  the  diodes  can  be  reﬂected  back  oﬀ  of  the  skin  to  detected

and  quantiﬁed  by  the  signal  detector.   The  degree  of  light  reﬂected  by  the  skin  depends  on

many various factors,  but importantly it depends on the volume and oxygenation of the blood

in  underlying  blood  vessels  and  capillaries  of  the  skin.  The  greater  the  volume  of  blood  the

more  the  signal  is  attenuated  and  therefore  the  lower  the  signal  measured  at  the  detector.

Since  the  volume  of  blood  in  the  vessel  varies  with  the  state  of  the  cardiovascular  cycle  it

is  possible  to  infer  the  patient’s  pulse  rate  from  the  detected  waveform.   While  the  signal

waveform  is  less  reliable  and  information-laden  than  an  electrocardiograph  waveform  it  does

10

have  the  advantage  of  being  non-invasive,  inexpensive,  highly  portable  and  easy  to  measure,

not  requiring  the  use  of  unruly  wired  electrodes.  We  recommend  the  following  comprehensive

review by J. Allen of the technical details and intricacies of photoplethysmography technology

for  the  interested  reader  [[52](#link_bookmark_164)].

The Fitbit wearable collects this raw PPG data and using Fitbit’s proprietary PurePulse algo-

rithm to infer the heart rate from the waveform data [[51](#link_bookmark_163)].  The exact details and functioning of

this  algorithm  is  not  disclosed  to  the  general  public  and  subject  to  change  (and  be  upgraded)

at  any  time.  Although  the  algorithm  is  not  infallible  and  not  endorsed  by  Fitbit  for  clinical

use,  Fitbit  devices  which  are  regularly  independently  reviewed  by  researchers  usually  score  as

one  of  the  most  accurate  commercial  activity  trackers  (generally  with  about  at  least  80-90%+

accuracy  on  heart  rate)  [[22](#link_bookmark_119), [24](#link_bookmark_134), [25](#link_bookmark_168), [38](#link_bookmark_169), [39](#link_bookmark_170), [53](#link_bookmark_171)].  Heart  rate  data  is  available  for  every  1  second

interval  if  the  device  is  set  in  workout  mode  (for  compatible  devices  like  the  Fitbit  Charge

HR),  and  every  5  seconds  otherwise  [[22](#link_bookmark_119)].

Once  the  raw  waveform  is  processed  the  heart  rate  data  is  uploaded  using  the  Bluetooth

wireless  communication  standard  to  a  paired  device  (typically  a  smartphone).   The  smart-

phone  then  relays  this  information  to  Fitbit  servers  where  it  is  stored  for  use  by  the  user,

by  Fitbit  and  by  authorized  applications.  This  is  done  automatically  and  seemlessly  without

user intervention so long as the devices (Fitbit,  Phone and Fitbit servers) are able to maintain

communication with each other.  The Fitbit is able to store data for a week without uploading

it  before  it  begins  to  aggregate  and  overwrite  old  data.

Fitbit  exposes  an  applications  programming  interface  (API)  which  can  be  used  by  authorized

parties  to  access  user  data  (with  the  requisite  user  permission)  on  Fitbit  servers.  Using  this

API  the  back-end  servers  that  power  part  of  the  Medly  telemonitoring  platform  are  able

to  retrieve  study  participant  data  from  the  Fitbit  servers.   Study  participant  activity  data

uploaded  to  Fitbit  servers  was  automatically  requested  by  the  Medly  servers  on  a  daily  basis.

Heart  rate  data  was  requested  only  for  every  60  second  interval  during  the  study  period  due

to  storage  constraints.

2.  Step  Count  [steps  per  minute]  -  recorded  using  commercially  available  activity-tracker  (Fitbit

Charge  HR,  Charge  2  or  Alta  HR)  continuously  throughout  the  day.

Fitbit  activity-trackers  measure  step  count  using  an  integrated  3-axis  accelerometer  [[54](#link_bookmark_173)].  Ac-

celerometers  convert  the  acceleration  they  experience  into  useful  electrical  signals.    3-axis

accelerometers  in  particular  measure  accelerations  in  all  3  orthogonal  cartesian  axes  and  so

can  record  movements  in  any  direction.   The  algorithms  that  interpret  the  various  acceler-

ation  signals  into  speciﬁc  human  motions  are  complicated  and  far  beyond  the  scope  of  this

work.   As  with  PurePulse  algorithm,  the  exact  algorithm  used  by  Fitbit  to  convert  raw  ac-

celerometer  data  into  step  count  is  proprietary  and  not  released  to  the  general  public  [[54](#link_bookmark_173)].

The  step  count  algorithm  is  also  known  not  to  be  100%  reliable,  although  it  continuously

gets  tuned,  upgrade  and  improved  by  Fitbit  [[55](#link_bookmark_175)].  As  with  the  heart  rate  data  independently

reviews  by  researchers  usually  score  Fitbit  devices  as  being  generally  high  accurate  in  their

daily measurement of step counts (again usually with at least 80-90%+ accuracy on step count

measurements)  [[24](#link_bookmark_134)–[26](#link_bookmark_135), [29](#link_bookmark_122), [38](#link_bookmark_169), [39](#link_bookmark_170), [56](#link_bookmark_181)]

As  with  the  heart  rate  data,  once  the  Step  Count  is  inferred  from  the  raw  accelerometer

data  the  Step  Counts  are  uploaded  using  the  Bluetooth  wireless  communication  standard  to

a  paired  device  (typically  a  smartphone).   The  smartphone  again  relays  this  information  to

11

Fitbit  servers  where  it  is  stored  for  use  by  the  user,  by  Fitbit  and  by  authorized  applications.

This  is  done  automatically  and  seemlessly  without  user  intervention  so  long  as  the  devices

(Fitbit,  Phone  and  Fitbit  servers)  are  able  to  maintain  communication  with  each  other.  The

Fitbit  is  also  able  to  store  this  data  for  a  week  without  uploading  it  before  it  begins  to

aggregate  and  overwrite  old  data.

Fitbit  also  exposes  the  applications  programming  interface  (API)  for  step  count  data  which

the  Medly  servers  use  to  request  the  data  on  a  daily  basis.  Step  count  data  will  also  only  be

requested  for  every  60  second  interval  (matching  the  heart  rate  data  resolution)  during  the

study  period  due  to  storage  constraints.

3.  Patient  Demographic/Meta  Data  -  recorded  when  prescribed  Medly  (carried  over  from  on-

boarding  process  before  this  study),  namely:

(a)  Sex  [Male  or  Female]

(b)  Age  [years]

(c)  Handedness  [left  or  right]

4.  Prescribed  Medications  -  recorded  at  on-boarding  and  updated  as  required  at  every  clinic

visit,  sourced  from  clinic  notes:

(a)  Beta  Blockers  [mean  daily  dose  in  mg]

All  of  the  aforementioned  predictors  are  stored  on  the  Medly  telemonitoring  platform  servers.

The  individual  data  elements  are  linked  to  a  unique  Medly  user  account  for  each  study  participant.

Data  stored  on  Medly  servers  in  this  format  is  then  easily  exportable  and  convertible  from  it’s

native  format  to  a  format  more  suitable  for  further  processing  using  an  analytic  software  of  choice

(e.g.  R).

**2.5     Sample  Size**

As  a  general  heuristic  (or  even  simply  a  frame  of  reference)  machine  learning  practitioners

generally  consider  data  sets  on  the  order  of  hundreds  of  samples  to  be  relatively  small  [[57](#link_bookmark_182)–[59](#link_bookmark_183)].  The

exact  size  of  a  data  set  required  to  properly  train  a  typical  Hidden  Markov  Model  (or  any  machine

learning  algorithm  in  general)  depends  on  a  number  of  diﬀerent  factors  including  the,  method

of  classiﬁcation,  complexity  of  the  classiﬁer,  separation  between  classes,  variance  and  presence  of

noise  in  the  data,  amongst  many  other  factors.   The  noisier,  the  more  complex  and  the  greater

the  variance  in  the  data,  typically  the  larger  the  dataset  required  to  achieve  good  performance.

There  is  no  upper  limit  for  how  much  data  should  be  used  for  training  but  there  point  at  which

increasing  input  data  begins  to  yield  diminishing  returns  in  improving  predictive  performance  [[59](#link_bookmark_183)].

The  exact  relationship  between  training  set  size  and  predictive  performance  for  an  algorithm  and

problem  in  question  is  often  shown  as  a  ’learning  curve’  graph  (which  plots  training  set  size  versus

prediction  error(s)).   To  the  best  of  the  author’s  knowledge  the  learning  curve  for  this  particular

application  (or  a  suﬃciently  analogous  application)  has  not  yet  been  determined.  However,  given

that  we  expect  that  the  data  collected  in  this  study  will  be  relatively  noisy  and  complex  we  expect

that  the  model  may  lean  towards  requiring  more  data  rather  than  less  data.  Therefore,  we  collect

as  much  data  as  was  available  in  order  to  not  prematurely  limit  the  power  nor  the  generalizability

of  the  algorithm  developed.

12

All  this  being  said,  given  a  present  recruitment  rate  of  approximately  1  patient  per  week  at  the

Heart  Function  Clinic  we  expect  to  be  able  to  accumulate  a  dataset  of  approximately  30  patients

over  the  30  week  period  from  the  beginning  of  December  2017  to  the  end  of  April  2018.  Based  on

a  previous  study  performed  at  the  Heart  Function  Clinic  several  years  ago  we  expect  a  distribution

of  approximately  40/15/40/5  for  patients  in  classes  II,  undetermined  II/III,  III  and  IV  respectively

-  so  approximately  12,  4  to  5,  12  and  1  to  2  patients  in  each  corresponding  class  [[60](#link_bookmark_185)].

**2.6     Missing  Data**

Due  to  the  nature  of  the  data  collection  methods  used  we  don’t  expect  to  have  a  lot  of  missing

data.    However,  despite  initial  expectations  missing  data  is  still  often  an  inevitable  reality  for

scientiﬁc  studies.  We  handle  missing  data  in  the  following  manner:

Missing Heart Rate and Step Count data is not explicitly reported by the activity tracker.  If the

device  is  not  being  worn  it  will  report  both  step  count  and  heart  rate  values  as  0  from  which  we  can

infer  that  the  data  is  missing.  If  step  count  is  zero  while  heart  rate  is  non-zero  we  can  infer  that  the

patient  is  simply  at  rest.  We  suspect  that  use  of  the  Fitbit  will  be  more  of  an  all-or-nothing  aﬀair

where  patients  put  on  the  device  at  the  beginning  of  the  day  and  keep  it  on  all  day  as  opposed  to

constantly using and removing the device.  As a result,  patients with less than 30 minutes of average

daily  ’wear  time’  during  each  analysis  period  will  be  ﬂagged  since  they  will  likely  have  insuﬃcient

data  for  each  day  for  the  system  to  perform  a  reliable  classiﬁcation.

Blood pressure data which always comes as a set of systolic/diastolic readings are imputed using

non-parametric  multiple  imputation  using  the  ’missForest’  R  package  (which  uses  a  random  forest

algorithm  to  predict  missing  values  and  makes  no  assumption  about  the  underlying  distribution  of

the data and supports both missing categorical and continuous variables) [[61](#link_bookmark_186),[62](#link_bookmark_187)].  Weight, symptom

questionnaire  responses,  heart  failure  treatments  to  date  and  prescribed  medications  are  dealt  with

using  the  same  approach.

Patients  without  any  speciﬁed  NYHA  class  (i.e.  missing  outcome  data)  are  ignored  completely,

however  gaps  in  NYHA  class  outcomes  are  imputed  using  the  same  approach  as  missing  predictors

(namely  blood  pressure,  weight,  etc.).

**2.7     Statistical  Analysis  Methods**

**2.7.1     Normalization**

Before  using  the  data  to  train  the  classiﬁcation  algorithm  the  predictors  will  be  normalized  to

values between 0 and 1 (after imputation of missing data as required).  The normalization procedure

for  each  of  the  predictors  is  as  follows:

1.  Heart  Rate  [beats  per  minute]  -  linear  scaling  using:  *min*= 0,  *max*= 208 *−−*0*.*7 *∗−age*(where

patient’s  age  is  measured  in  years)  [[63](#link_bookmark_188)]

2.  Step  Count  [steps  per  minute]  -  linear  scaling  using:  *min*= 0,  *max*= 300

(a)  Sex  [Male  or  Female]:  *Male*= 0,  *Female*= 1

(b)  Age  [years]-  linear  scaling:  *min*= 18,  *max*= 150

(c)  Handedness  [left  or  right]:  *Lef t*= 0,  *Right*= 1

3.  Prescribed  Medications  [mean  daily  dose  in  mg]  -

13

(a)  Beta    Blockers    -    linear    scaling:      Acebutolol    *min*=     0,    *max*=     1200

[[64](#link_bookmark_189)]

&   Atenolol   *min*=    0,    *max*=    300   &   Bisprolol   *min*=    0,    *max*=    20

& Carvedilol *min*= 0, *max*= 80 & Labetalol *min*= 0, *max*= 2400 &Metaprolol *min*=

0,       *max*=        450      &      Nadolol      *min*=

0,       *max*=

320

&  Pindolol  *min*= 0,  *max*= 60  &  Propranolol  *min*= 0,  *max*= 640  &  Timolol  *min*= 0,

*max*= 60  [[65](#link_bookmark_190), [66](#link_bookmark_191)]

**2.7.2     Classiﬁer  (Model)  Development**

The  algorithm  we  develop  to  classify  patients  into  their  corresponding  NYHA  class  will  be

designed  to  perform  either  binary  classiﬁcation  of  class  II  and  II  patients,  or,  dependant  on  our

ability  to  capture  class  IV  patients  for  analysis,  multi-class  classiﬁcation  of  class  II,  III  and  IV  .

The  following  section  outlines  the  details  of  the  development  in  both  cases  (which  are  very  similar

except  for  the  addition  of  a  third  class).

In  either  case  we  propose  initially  exploring  three  diﬀerent  levels  of  classiﬁers  complexity,  i.e.

that take into account a varying subset of the available predictors outlined in section [2.4](#Predictors), speciﬁcally

the  three  following  subsets:

1.  step  count  only

2.  step  count  +  heart  rate

3.  step  count  +  heart  rate  +  prescribed  medications

Looking  at  three  diﬀerent  levels  of  classiﬁer  complexity  will  be  informative  since  while  the  Medly

telemonitoring platform supports the highest level of complexity speciﬁed above, certain researchers

and  clinicians  may  only  have  integrated  access  to  prescribed  medications  along  with  activity  and

heart  rate  monitoring  and  furthermore  possibly  have  ready  access  to  ﬁtness  monitors  that  monitor

both  activity  and  heart  rate.  It  may  be  that  using  too  many  variables  would  make  our  model  too

complex  rendering  it  even  more  prone  to  over  ﬁtting  given  the  available  data.

Additionally,  as  a  reminder  both  for  training  and  classiﬁcation,  we  segment  each  patient  data

stream  into  two-week  rolling  windows  of  data  so  that  the  algorithm  developed  will  be  designed  to

classify  patients  using  rolling-two  week  chunks  of  data  instead  of  requiring  a  several  month  long

window.

All  this  said,  returning  to  the  task  at  hand;  to  perform  the  classiﬁcation  we  create  and  train

multivariate  Hidden  Markov  Models  for  each  of  the  target  risk  groups  (NYHA  class  II,  III,  IV)

and  so  before  we  begin  however  it  is  probably  important  that  we  review  some  basics  about  Hidden

Markov  Models.

**Basics of Markov Models (Hidden or Otherwise)**Markov Models are probabilistic state ma-

chines where the transitions between states are executed randomly according to pre-speciﬁed transi-

tion probabilities between states [[67](#link_bookmark_193)–[71](#link_bookmark_194)].  MarkovModels are used to modelMarkov chains/processes

which  are  stochastic  (i.e.  random)  processes  that  satisfy  the  Markovian  property:  that  is,  the  tran-

sitions  from  a  given  state  in  the  chain  to  the  next  immediate  state  (and  by  extension  all  future

states)  must  be  dependant  solely  on  the  current  state  of  the  model  [[67](#link_bookmark_193)–[72](#link_bookmark_196)].  They  must  not  depend

on  the  path  taken  to  arrive  to  that  state,  i.e.  it  cannot  depend  on  any  previous  states  in  which  the

system  has  existed.  The  Markovian  property  is  alternatively  known  as  the  ’memoryless’  property;

essentially that the Markov process or markov chain has no memory of the past [[67](#link_bookmark_193)–[72](#link_bookmark_196)].  And so, the

14

Figure  1:  Markov  Model

transition  probabilities  along  with  the  number  of  states  form  the  fundamental  model  parameters

that  can  be  used  to  uniquely  describe  the  Markov  Model.   Where  relevant  a  Markov  Model  may

also  have  initial  starting  parameters  which  dictate  the  likelihood  associated  with  the  Markov  Model

starting  in  each  possible  state  (e.g.  10%  chance  to  start  in  State  S1,  20%  chance  to  start  in  State

S2  and  so  on)  [[67](#link_bookmark_193)–[72](#link_bookmark_196)].

In  many  Markov  Models  (and  in  every  Hidden  Markov  Model)  there  are  also  an  associated  set

of  possible  observations  that  are  linked  to  each  state,  i.e.   that  can  possibly  be  output  by  each

state.   For  example  assume  a  Markov  Model  that  models  weather  outside  an  oﬃce  with  possible

states  S1  =  Sunny,  S2  =  Cloudy  and  S3  =  Rainy  with  associated  transition  probabilities  between

each  state  [[68](#link_bookmark_201)].   The  observations  associated  with  each  state  might  be  the  clothing  that  a  given

person  in  a  stream  of  passers-by  are  wearing,  say  a  shirt,  a  sweater  or  a  rainjacket  [[68](#link_bookmark_201)].   It  is

possible  that  a  person  might  be  wearing  any  of  these  types  of  clothing  in  any  given  type  of  weather

but  it  is  likely  the  case  that  the  likelihood  of  observing  each  clothing  type  will  diﬀer  based  on

the  underlying  weather  state;  for  example  rainjackets  are  probably  more  likely  to  be  observed  in

rainy  weather  than  in  sunny  weather  [[68](#link_bookmark_201)].  These  probabilities  are  termed  observation  probabilities

and  link  the  states  in  the  Markov  Model  to  the  observations  that  are  measured  as  outputs  of  the

Markov  Model.  These  observations  could  be  speech  phonemes,  written  characters  of  the  alphabet,

or genome sequences [[71](#link_bookmark_194),[73](#link_bookmark_205)].  Figure [1 s](#link_bookmark_206)

example  (including  the  starting,  transition  and  observation  probabilities).

The  appropriately  named  Hidden  Markov  Models  (HMM)  are  simply  Markov  Models  where  the

underlying  states  are  hidden  -  i.e.  cannot  directly  be  observable  [[67](#link_bookmark_193), [69](#link_bookmark_208)–[72](#link_bookmark_196)].  Speciﬁcally,  while  we

may know the number of states in the model, the transition probabilities of the model, we don’t know

the  exact  state  the  system  is  in  or  the  sequence  of  states  it  has  been  through  [[67](#link_bookmark_193),[69](#link_bookmark_208)–[72](#link_bookmark_196)].  However,

15

given the observation probabilities it is possible for us to try and infer the current state of the Model,

including the sequence of states that a particular Markov Model went through [[67](#link_bookmark_193),[69](#link_bookmark_208)–[72](#link_bookmark_196),[74](#link_bookmark_216)].  This is

because Hidden Markov Models,  as with Markov Models,  operate under the Markovian assumption:

they assume that the process they model adhere to the Markovian property [[67](#link_bookmark_193),[68](#link_bookmark_201),[71](#link_bookmark_194),[72](#link_bookmark_196)].  However,

it  has  been  found  that  Hidden  Markov  Model  are  able  in  certain  cases  to  fairly  successfully  model

processes that violate this Markovian assumption such as speech recognition and gesture recognition

[[71](#link_bookmark_194), [73](#link_bookmark_205), [75](#link_bookmark_223)].  Part  of  contributions  of  this  study  will  be  to  determine  if  Hidden  Markov  Models  are

suitable for modelling patient activity and heart rate data both of which likely violate the Markovian

assumption  ’demanded’  of  Hidden  Markov  Models.

**Semi-Markov  Model**The  violation  of  the  pure  Markovian  assumption  leads  us  to  a  variation

on  Hidden  Markov  Models:  Hidden  Semi-Markov  Models  (HSMM)  [[70](#link_bookmark_224)].  HSMMs  are  HMMs  that

formally  relax  the  ’Markovian’  assumption  of  the  model  by  permitting  the  model  to  speciﬁcally

retain  the  memory  of  how  long  it  has  been  in  a  certain  state  (typically  in  order  to  force  the  model

to not exist in a state for more than a desired time) [[70](#link_bookmark_224)].  As such, HSMMs require that an additional

set  of  parameters  be  deﬁned:  the  sojourn  distribution  of  each  state  [[70](#link_bookmark_224)].  That  is,  the  distribution  of

expected  mean  waiting  times  in  each  given  state.  These  waiting  times  can  follow  any  distribution

desired  -  normal,  geometric,  gamma,  etc.  -  or  appropriate  for  the  problem  at  hand  [[70](#link_bookmark_224)].

In  our  case,  since  it  is  probably  unreasonable  to  assume  that  there  no  some  time-dependence

in  state  changes  due  to  the  dynamic  nature  of  human  exercise  and  activity  (e.g.   people  who  are

performing  high-intensity  activity  are  less  likely  to  continue  as  time  goes  by  since  they  get  tired)

we  propose  also  training  equivalent  multivariate  hidden  semi-Markov  models  for  each  of  the  target

risk  groups  to  explore  the  eﬀect  of  relaxing  the  Markovian  assumption  (or  time-independence)  of

pure  Markov  models  if  the  regular  pure  Markov  models  prove  to  be  unsuitable  for  our  study.

**Parameters  Required  for  Hidden  Markov  &  Semi-Markov  Models**To  summarize,  the

complete  set  of  parameters  that  must  be  determined  for  a  Hidden  Markov  Model  are  as  follows:

1.  the  number  of  states  in  the  model

2.  the  starting  probabilities

3.  the  transition  probabilities

4.  the  (observation)  emission  probabilities

For  Hidden  Semi-Markov  Models,  we  add  the  additional  parameter  of  the  individual  state  sojourn

distributions.

**Determining  Markov  Model  Parameters**Determining  the  single  best  or  most  optimal  Hid-

den  Markov  Model  parametrization  for  given  data  stream  is  unfortunately,  an  intractable  prob-

lem  [[67](#link_bookmark_193), [69](#link_bookmark_208), [71](#link_bookmark_194)].   That  being  said,  there  is  a  known  algorithm  for  eﬃciently  computing  the  most

likely  locally  optimal  parametrization,  the  maximum  likelihood  estimation,  for  a  stream.  Generally

speaking  the  speciﬁc  sub-class  of  algorithms  used  to  solve  this  problem  in  the  markov  model  space

are  known  as  expectation-maximization  (EM)  algorithm  [[67](#link_bookmark_193), [69](#link_bookmark_208), [71](#link_bookmark_194)].   One  of  the  most  common

EM  algorithm  implementations  used  for  Hidden  Markov  Model  training  is  the  Baum-Welch  algo-

rithm  [[67](#link_bookmark_193), [69](#link_bookmark_208), [71](#link_bookmark_194), [76](#link_bookmark_237)].  Another  common  algorithm  used  to  approximate  EM  is  the  Viterbi  *training*

16

algorithm  (N.B.  not  the  Viterbi  algorithm)  which  can  yield  less  accurate  models  than  the  Baum-

*s*

*t*

*t*

*t*

Welch  algorithm  but  is  usually  much  less  computationally  intensive  [[76](#link_bookmark_237), [77](#link_bookmark_239)].   We  eschew  further

discussion  of  the  implementation  details  of  either  of  these  algorithms  since  the  availability  of  pre-

programmed  libraries  implementing  these  algorithms  makes  it  unnecessary  for  any  given  arbitrary

reader  to  have  the  in-depth  knowledge  required  to  implement  the  algorithms  and  because  there  are

many  excellent  sources  available  that  explain  the  ﬁne  details  of  algorithm  probably  much  better

than  this  author  could  [[67](#link_bookmark_193), [69](#link_bookmark_208), [71](#link_bookmark_194), [77](#link_bookmark_239)].   In  any  case  none  of  these  algorithms  is  able  to  determine

all  of  the  parameters  by  itself.  Some  of  the  parameters  must  be  provided  as  ’initial  conditions’  for

the  algorithm  to  execute.  Typically  these  are  the  emission  probabilities,  the  starting  probabilities,

the  sojourn  distributions  (and  sometimes  even  initial  transition  probabilities).   Depending  on  the

library used it may try to make an educated guess for starting points or leave the ’initial conditions’

to  be  speciﬁed  solely  by  the  author.   It  is  possible  (and  encouraged)  to  try  various  combinations

of  parameters  to  determine  the  most  eﬀective  set  -  in  fact  more  fully  featured  software  libraries

will  also  sometimes  oﬀer  to  do  this  automatically,  although  it  is  ultimately  up  to  the  researcher  to

determine  appropriate  ’initial  conditions.’

In  our  case  we  use  the  ’mhsmm’  package  for  the  R  software  platform  [[70](#link_bookmark_224),[78](#link_bookmark_245)]  which  requires  the

user to provide the number of desired states, the emission probabilities (which are assumed to remain

ﬁxed) as well as an initial starting point for the state probabilities and transition probabilities, which

the  algorithm  then  adjusts  as  it  searches  for  a  local  optimum.

**Proposed  Initial  Markov  Model  Parameters**For  this  study  we  anticipate  to  start  with  the

following  initial  conditions:

1.  **States**:  We  begin  with  3  initial  states  and  then  sweep  from  3  to  6-8  states  depending  on  the

available  computational  power.  The  computational  power  limit  is  important  since  computa-

tional  complexity  for  training  Hidden  Markov  Models  increases  to  the  square  of  the  number

of  states  as  a  result  of  the  states  being  interconnected.   That  is,  with  3  states  there  are  9

possible  transitions  between  states  which  must  be  solved.  Doubling  the  number  of  states  to

6  causes  a  quadrupling  of  the  number  of  possible  transitions  to  36.  At  8  states  there  are  64

possible  transitions,  almost  double  that  of  the  6  state  case  which  may  make  such  an  HMM

too  unwieldly  for  us  to  train  given  available  resources.

2.  **Starting  State  Probabilities**:  We  set  a  probability  of  1  (i.e.   100%  likelihood)  for  one  of

the  states  and  0  for  all  the  others.  We  anticipate  that  this  will  be  a  good  starting  point  based

on  initial  exploration  of  the  data  that  seemed  to  indicate  that  patients  spend  most  of  their

time  in  a  non-active  state.   In  other  words,  at  any  given  moment  of  we  were  to  look  at  the

time  stream  it  is  most  likely  that  a  patient  will  be  in  the  non-active  state  as  opposed  to  any

other  state  so  we  reﬂect  this  in  the  starting  probabilities.

3.  **Transition  Probabilities**:   Given  a  lack  of  evidence  to  the  contrary  we  make  the  simple

assumption  that  the  transition  probabilities  are  uniform  and  so  set  each  to  be  the  reciprocal

of  the  number  of  state  transitions.  In  other  words,  if  the  we  let  *n*be  the  number  of  possible

states,  and  *n*be  the  number  of  possible  state  transitions,  the  initial  transition  probabilities

1

1

*P*are  each  *P*=

*n*

*s*

2

=

*n*

*t*

.

4.  **Emission  Probabilities**:  Based  on  previous  analysis  of  the  data  the  minute-by-minute  step

counts  appear  to  be  approximately  gamma  distributed  with  peaks  (or  sub-peaks)  clusters.

We  therefore  propose  setting  the  emission  probabilities  to  be  gamma  distributed  such  that

17

the  distributions  center  around  prominent  peaks  visually  identiﬁed  from  the  data:   e.g.   20

steps/min,  80  steps/min,  100  steps/min  for  a  3  state  HMM.  For  heart  rate  emission  probabil-

ities  we  similarly  propose  using  normal  distributions  initially  centred  around  the  traditional

heart  rate  zones  which  for  a  normal  healthy  human  are  deﬁned  as  follows  [[51](#link_bookmark_163)]:

(a)  resting:  60  bpm

(b)  above  resting  but  *<*50%  of  max,  slightly  elevated:  89  bpm

(c)  50  to  69%  of  max,  low  intensity  exercise:  110  bpm

(d)  70  to  84%  of  max,  medium  intensity  exercise:  135  bpm

(e)  85%  to  100%  of  max,  high  intensity  exercise:  152  bpm

(f)  *≥*100%  of  max,  maximum  intensity  exercise:  220  bpm  -  age

5.  **Sojourn  Distributions**:  For  exploration  of  Hidden  semi-Markov  Models  the  propose  using

an  identical  gamma  distribution  for  each  state  to  model  the  mean  waiting  time.  The  gamma

distributions  would  be  centered  around  a  certain  period  of  time,  sweeping  that  period  of  time

in  1hr  intervals  from  1  to  12  hrs,  under  the  assumption  that  our  patients  person,  despite

their  condition,  will  likely  have  to  get  up  (or  attempt  to  get  up  -  thereby  registering  even

momentary  activity)  once  a  day  to  attend  to  essential  needs.

Now that we have reviewed the essentials of Markov Models and how to compute the parameters

of a Hidden Markov Model in general (including the proposed initial parametrizations of each of the

models)  we  proceed  to  discuss  how  we  propose  to  use  Hidden  Markov  models  to  classify  patients.

**Classiﬁcation**A  previously  mentioned  to  perform  the  classiﬁcation  we  propose  creating  and

train  multivariate  Hidden  Markov  Models  for  each  of  the  target  risk  groups  (NYHA  class  II,  III,

IV).  We  can  collect  the  Hidden  Markov  Models  that  were  generated  into  a  sets  with  a  Hidden

Markov  Model  for  each  risk  group,  i.e.   for  a  binary  classiﬁer  the  sets  would  consist  of  1  HMM

trained  using  NYHA  class  II  patients,  and  1  using  NYHA  class  III  patients.   By  extension  for

the  3  class  multi-class  classiﬁer  the  sets  would  additional  contain  a  HMM  trained  using  NYHA

class  IV  patients.   Classiﬁcation  can  then  be  performed  by  evaluating  the  likelyhood  that  a  given

(unclassiﬁed)  patient’s  sequence  (i.e.  activity,  heart  rate  data  stream)  was  generated  from  each  of

the  corresponding  HMMs  in  a  set  -  the  evaluation  of  which  can  performed  using  either  the  forward

or backward algorithm detailed in any of these referenced works, and whose functionality is included

in  most  HMM  programming  libraries  [[67](#link_bookmark_193), [69](#link_bookmark_208)–[72](#link_bookmark_196)].  The  predicted  classiﬁcation  of  the  patient  would

then  correspond  to  the  class  of  the  model  with  the  highest  likelihood  of  having  ’generated’  that

data  stream.  This  process  is  detailed  graphically  in  Figure  2.  In  an  ideal  system  we  would  select

the  single  set  of  Hidden  Markov  Models  that  would  be  able  to  best  classify  patients  into  their

appropriate  class  -  how  we  go  about  selecting  this  set  is  detailed  later  in  this  section  as  part  of  the

Hidden  Markov  Model  training.  In  general  though,  and  in  particular  within  the  context  of  a  larger

system  in  which  patients  might  be  stratiﬁed  by  other  variables,  such  as  sex  or  the  use  of  a  certain

medications (e.g.  beta-blockers),  one could feasible also train an ideal Hidden Markov Model set for

each  patient  sub-group  or  stratiﬁcation.  This  process  by  which  this  is  done  for  a  group  of  patients,

collected  together  as  the  *Data*block  is  detailed  graphically  in  Figure  3.  The  output  of  this  process

can  be  constructed  as  a  table  where  each  patient  is  listed  along  with  it’s  corresponding  predicted

class  -  a  *Patient  Class  Table*.

18

Figure  2:  C2:  Classiﬁer  -  Classify  Using  Model  Set

19

Figure  3:  C0&C1  Classiﬁer  -  Generalized  System

20

**Training:   Preparing  &  Preprocessing  of  Data**Having  outlined  how  the  proposed  NYHA

classifer  is  supposed  to  work  we  turn  to  the  task  of  training  a  system,  that  is,  determining  the

ideal  Hidden  Markov  Model  set  for  a  provided  set  of  input  patients.  As  a  ﬁrst  step  to  this  training

however it is important that we ﬁrst preprocess and clean the *raw data*that we will use as input.  The

preprocessing  procedure  is  outlined  in  Figure  4.  First  the  raw  data  is  passed  through  an  *automated*

*cleaner*which  is  an  script  that  ’cleans’  and  standardizes  the  data  before  use,  including  removing

or  ﬁlling  in  missing  variables  as  required,  standardizing  keywords  so  that  they  all  follow  the  same

format (e.g.  NYHA Class II is cleaned and converted so that it is always internally represented as ’II’

instead of ’ii’,’2’,’two’ or any other variants) or removing characters, data elements or otherwise that

are  not  needed  or  might  even  cause  problems  with  any  other  downstream  processing  elements.  This

automated cleaner is updated as required on each pass (by the author) to include any common items

that  have  been  detected  during  manual  cleaning  and  veriﬁcation  of  the  data  so  that  the  system  can

more easily be transferred to a live clinical environment and to improve the ability of the *automated*

*cleaner*to  completely  clean  the  data.   After  passing  through  the  *automated  cleaner*and  before

*Manual  Cleaning  and  Veriﬁcation*the  data  is  passed  through  a  preprocessing  and  formatting  step

where  the  data  is  converted  from  the  raw  input  data  type  (for  example  tab-deliminated  string)  into

a  format  more  amenable  for  the  processing  environment  (for  example  an  R  dataframe,  vector  or

matrix).  As  previous  mentioned,  after  this  step  the  data  is  manually  veriﬁed  to  catch  any  obvious

errors.  The  output  of  this  last  veriﬁcation  process  is  the  ’Processed  Data’.

Due  to  the  particular  circumstances  of  our  study,  namely  that  the  data  set  collected  is  too

small  to  perform  practical  machine  learning  which  requires  a  large  amount  of  data  this  *processed*

*data*undergoes  another  transformation  before  it  is  used  to  train  our  machine  classiﬁer.  This  is  the

microsimulation  step.   Microsimulation  is  useful  in  the  case  where  a  particular  machine  learning

algorithm  needs  to  see  a  minimum  ’critical  mass’  of  data  in  order  to  be  able  to  begin  to  successfully

classify  inputs.  Speciﬁcally  microsimulation  allows  us  to  generate  an  arbitrary  amount  of  simulated

’fake’  data  from  an  original  dataset  based  on  pre-identiﬁed  statistical  properties  of  the  original

dataset  (with  a  little  added  noise  for  variance).    The  idea  is  that  since  this  simulated  data  is

(statistically)  representative  of  the  original  data  the  machine  learning  algorithm  will  pick  up  on  the

same traits as it would have from an original larger dataset but without requiring that larger dataset

in  the  ﬁrst  place.  There  are  of  course  limitations  to  the  inferences  we  can  make  on  the  classiﬁcation

accuracy  of  a  machine  learning  model  trained  using  microsimulated  data,  but  it  is  a  useful  tool  for

demonstrating  proof  of  concept.  In  Figure  4  microsimulation  is  shown  as  being  performed  in  two

parts,  ﬁrst  the  statistical  properties  of  the  *processed  data*analysed  and  the  results  of  this  analysis

are  then  used  to  *microsimulate*(i.e.  generate)  the  *µSimulated  Data  For  Training*.  This  *µSimulated*

*Data  For  Training*forms  the  dataset  used  for  the  rest  of  the  development  of  the  models.  Although

for  the  purposes  of  this  explanation  we  do  not  stratify  the  patients,  the  process  of  stratiﬁcation,

if  required,  would  be  performed  after  generating  the  processed  data  and  before  performing  the

microsimulation  (including  both  the  analysis  of  statistical  properties  and  the  *microsimluate*steps).

Stratifying  patients  at  this  point  allows  us  to  generate  a  corresponding  dataset  of  *µSimulated  Data*

*For  Training*for  each  patient  strata  which  we  can  separately  feed  as  the  input  for  the  rest  of  the

training  process  until  we  have  generated  the  corresponding  Hidden  Markov  Model  set  for  each  of

the  strata.

21

Figure  4:  P0:  Preprocessing  Data

22

**Training:  Model  Selection**After  converting  the  raw  data  into  clean  and  usable  format  we  can

proceed  to  the  actually  generation,  training  and  selection  of  our  candidate  Hidden  Markov  Models.

Overall  the  general  process,  which  is  shown  in  Figure  5  is  as  follows:

1.  Generate  potential  models,  which  we  store  as  a  *Potential  Model  List*.

2.  Apply  each  generated  candidate  model  to  each  patient  in  the  *µSimulated  Data  For  Training*

provided as input.  We do this to evaluate the probability that each candidate model generated

the corresponding patient data stream.  We keep these probabilities in a table which links each

model  to  the  probability  it  generates  for  each  patient;  this  is  the  *Model->Patient  Probability*

*Table*.

3.  Based  on  the  generated  *Model->Patient  Probability  Table*and  the  true  class  which  we  can

extract  from  the  *µSimulated  Data  For  Training*provided  as  input  we  can  evaluate  determine

the  best  model  set  of  the  candidate  sets  generated.

4.  Lastly  we  return  the  best  model  identiﬁed  in  the  previous  step.

**Training:   Model  Generation**The  process  of  generating  multiple  models  can  be  divided,  at

least logically, into two seperate parts.  The ﬁrst is that of generating a models for each of the classes.

The second involves generating diﬀerent models within each class group using diﬀerent initial HMM

parameters  with  the  goal  of  trying  to  ﬁnd  a  parametrization  that  is  closest  to  the  representing  the

true underlying ’model’ for that class group (i.e.  the globally optimal parametrization as opposed to

simply  a  locally  optimal  parametrization).  Both  of  these  are  faily  simple  and  the  combined  process

is  detailed  in  Figure  6.

The  ﬁrst  part,  generation  for  each  class  can  be  accomplished  by  simply  selecting  all  the  patients

that  belong  to  given  class  from  the  input  *µSimulated  Data  For  Training*and  using  these  as  the

training  data  for  the  model  training  function  *hmm.ﬁt*of  our  HMM  library  for  R:  ’mhsmm’.  This

training  function  outputs  a  potential  model  which  we  can  add  to  our  *Potential  Model  List*which

we  will  return  at  the  end  of  the  generation  process.

The  second  part,  generating  diﬀerent  parametrizations,  is  combined  as  part  of  the  class  by

class  generation  and  simply  involves  updating  the  initial  parameters  that  form  the  second  half  of

the  required  input  for  the  *hmm.ﬁt*training  function  until  we  have  swept  through  all  the  desired

parameter  variations.

**Training:  Extracting  the  Ideal  Model  Set**The  process  of  extracting  the  ideal,  or  single  best

model  set  from  the  generated  *Potential  Model  List*is  also  fairly  simple  and  is  detailed  in  Figure  7.

We  can  simply  iterate  through  every  possible  model  set,  i.e.  each  possible  combination  of  NYHA

class  II,  III  and  IV  models  (for  the  3  class  multi-class  classiﬁer)  and  evaluate  the  performance  score

of  each  model  set  by  comparing  the  *True  Classes  for  Patients*against  the  *Predicted  Classes  for*

*Patients*generated by the classiﬁer resulting from that set of HMMs (recall Figure 2).  We can store

these  scores  as  a  *Model  Set->Score  Table*for  future  review  or  as  detailed  in  Figure  5  simply  return

the  single  best  model  combination.   The  process  in  Figure  7  outlines  the  storage  method  since  it

is  most  useful  for  a  researcher,  who  may  wish  to  review  the  score  results  to  perform  additional

analyses  on  the  performance  scores.

23

**Training:  Model  Evaluation**The  process  by  which  the  performance  scores  of  each  model  can

be  evaluated  is  visually  detailed  in  Figure  8.  This  particular  process  however  is  described  in  detail

in  section  [2.7.3](#Classifier_(Model)_Performance_Assessment)  [of  t](#Classifier_(Model)_Performance_Assessment)

24

Figure  5:  T0:  Training  -  Model  Selection

25

Figure  6:  T1:  Training  -  Potential  Model  Generation

26

Figure  7:  T2:  Training  -  Extraction/Selection  of  Ideal  Model  Set

27

Figure  8:  T3:  Training  -  Model  Evaluation

28

————

*x*represent  those  patients  predicted  by  the  test  as  belonging  to  this

*x*).   In  other,  words  sensitivity  measures  how

ˆ

ˆ

*c*

*c*

*c     c*

*c*

*c*

*c*

*c*

Algorithm

Classiﬁcation

**Summary**In summary, we reviewed the basics of Markov Models including variations on the such

models  where  the  underlying  assumptions  are  relaxed.  We  also  outlined  the  method  by  which  we

develop a binary or 3 class multi-class classiﬁer that uses sets of multivariate Hidden Markov Models

to  classify  patients  into  their  corresponding  NYHA  class  within  the  subset  of  classes  supported  by

the  classiﬁer.   We  do  this  by  generating  several  possible  sets  of  Hidden  Markov  Models  for  each

corresponding  class  and  selecting  the  combination  of  models  amongst  those  generated  that  produce

the  most  performant  classiﬁer  according  to  the  performance  metrics  outlined  in  the  next  section

(section  [2.7.3](#Classifier_(Model)_Performance_Assessment))  although  we  provide  a  brief  visual  summary  of  the  evaluation  procedure.

**2.7.3     Classiﬁer  (Model)  Performance  Assessment**

**Confusion  Matrix**After development of an classiﬁcation algorithm it will be important to assess

how  well  the  classiﬁer  is  able  to  correctly  assess  patient  NYHA  class.  Since  performance  evaluation

measures  diﬀer  for  binary  and  multi-class  classiﬁers  this  section  discuss  both.  Measures  of  classiﬁ-

cation  accuracy  for  a  classiﬁcation  test  or  algorithm  (regardless  of  the  number  of  supported  classes)

can  be  reported  in  a  confusion  matrix,  an  example  of  which  is  shown  below  [[79](#link_bookmark_263)–[82](#link_bookmark_264)].

Clinical  Classiﬁcation

II

III

IV

Predictive  Value:

II

*a*

*b*

*c*

*a*

*a*+*b*+*c*

III

*d*

*e*

*f*

*e*

*d*+*e*+*f*

IV

*g*

*h*

*i*

*i*

*g*+*h*+*i*

Sensitivity:

*a*

*a*+*d*+*g*

*e*

*b*+*e*+*h*

*i*

*c*+*f*+*i*

Accuracy:

*e*+*f*+*h*+*i*

*d*+*f*+*g*+*i*

*d*+*e*+*g*+*h*

*a*+*e*+*i*

Speciﬁcity:

*b*+*c*+*e*+*f*+*h*+*i*

*a*+*c*+*d*+*f*+*g*+*i*

*a*+*b*+*d*+*e*+*g*+*h*

*N*=

P

*allcells*

**Standard  Performance  Measures**From  the  cells  of  a  confusion  matrix  we  can,  amongst  other

measures,  determine  the:

a)  *sensitivity*,  *recall*,  or  *true  positive  rate*of  a  test:  the  proportion  of  patients  correctly  identiﬁed

as  belonging  to  the  particular  test  class  e.g.   class  II  [[80](#link_bookmark_265), [82](#link_bookmark_264)–[86](#link_bookmark_267)].   Phrased  mathematically,

letting  *x*represent  those  patients  belonging  to  a  particular  class  *c*(e.g.   class  II)  from  the

set  of  all  classes  *C*and   ˆ

particular  class  then  the  sensitivity  is  *P*( ˆ *|x*

reliably  the  test  identiﬁes  patients  belonging  to  a  particular  class  by  labelling  them  with  that

class.

b)  *speciﬁcity*,  or  *true  negative  rate*of  a  test:  the  proportion  of  patients  correctly  identiﬁed  as  not

belonging  to  the  particular  test  class  e.g.   not  class  II  [[80](#link_bookmark_265), [82](#link_bookmark_264)–[86](#link_bookmark_267)].   Phrased  mathematically,

letting  !*x*represent  those  patients  not  belonging  to  a  particular  class  *c*(e.g.   not  class  II)

and  !*x*represent  those  patients  predicted  by  the  test  as  not  belonging  to  class  *c*then  the

sensitivity  is  *P*(!*x  |*!*x*).  In  other  words,  speciﬁcity  measures  how  reliably  the  test  identiﬁes

patients  not  belonging  to  a  given  class  by  labelling  with  a  class  other  than  that  given  class.

29

c)  *fall-out*,  or  *false  positive  rate*of  a  test:   the  proportion  of  patients  positively  identiﬁed  by

*c     c*

*x*=  *x*).   In  other  words,  the  accuracy  measures  how

*x*

*c*

*x*).  In other words, the positive predictive

S

*x*

*c*

*c*

*c*

*c*

1

1

*recall*

the  test  as  belonging  to  a  particular  class  but  actually  belonging  to  a  diﬀerent  test  class  e.g.

identiﬁed as class II but actually class I [[80](#link_bookmark_265),[82](#link_bookmark_264),[84](#link_bookmark_273)–[86](#link_bookmark_267)].  Phrased mathematically using the same

deﬁnitions as for the previous measures then the false positive rate is *P*( ˆ *|*!*x*

*c*

) = 1*−speciﬁcity*.

In  other  words,  the  false  positive  rate  identiﬁes  how  over-eager  the  test  is  at  labelling  patients

as  belonging  to  a  particular  class  in  an  eﬀort  to  not  miss  patients  that  actually  belong  to  that

class  -  it  is  the  ’fall-out’  or  ’collateral  damage’  in  the  ’aﬃrmative’  labelling.

d)  *precision*or  *positive  predictive  value*of  a  test:  the  proportion  of  patients  correctly  classiﬁed

as  belonging  to  a  particular  test  class  amongst  all  those  identiﬁed  by  the  test  as  belonging

to  that  class  [[82](#link_bookmark_264)–[86](#link_bookmark_267)].  Phrased  mathematically  using  the  same  deﬁnitions  as  for  the  previous

measures then the positive predictive value is *P*(*x  |−*ˆ

value  measures  how  likely  a  patient  is,  for  each  given  class,  to  actually  belong  the  class  it  has

been  labelled  as.

e)  *accuracy*of  a  test:  the  proportion  of  patients  correctly  classiﬁed  by  the  test  into  their  true

class  [[80](#link_bookmark_265), [82](#link_bookmark_264), [84](#link_bookmark_273)–[86](#link_bookmark_267)].   Phrased  mathematically  using  the  same  deﬁnitions  as  for  the  previous

measures  the  accuracy  of  the  test  is     *P*( ˆ  = *x*)  for  all  *c ∈−C*which  in  the  case  of  mutually

exclusive  classes  simpliﬁes  to

P

*P*( ˆ

*c∈C*

well  a  test  correctly  classiﬁes  patients  into  their  respective  classes.   Unfortunately,  the  way

accuracy  is  calculated  makes  it  highly  sensitive  to  the  underlying  base  rate  of  occurrence  of

each  of  the  classes,  namely  it  preferentially  weighs  classes  that  occur  more  frequently  in  the

underlying  population  often  limiting  it’s  usefulness  as  a  summary  metric.

f)  *F  score*,  *F    score*,  or  *F-measure*of  a  test:   an  alternative  to  accuracy  that  is  less  sensitive

to  the  underlying  base  rate  of  occurrence  for  each  of  the  classes.   It  is  the  harmonic  mean

of  *precision*and  *recall*for  that  class  [[80](#link_bookmark_265), [82](#link_bookmark_264), [84](#link_bookmark_273)–[86](#link_bookmark_267)].  Phrased  mathematically  using  the  same

deﬁnitions  as  for  the  previous  measures  the  F  score  of  the  test  is:

*F*= 2 *∗−*(

1

+

1

*precision*

)

*−*1

**Receiver Operating Characteristic & Area Under Curve**To aid in comparing performance,

many  of  the  above  test  measures  can  be  visualized  in  various  ways.  One  of  the  most  useful  is  the

Receiver  Operating  Characteristic  (ROC)  curve  which  shows  the  trade  oﬀ  between  the  *sensitivity*

and  *fall-out*of  a  test  [[79](#link_bookmark_263)–[83](#link_bookmark_286), [87](#link_bookmark_287)].  We  can  generate  the  ROC  curve  for  each  risk  group  by  plotting

the  corresponding  risk  group  *false  positive  rate*vs  *true  positive  rate*.   Alternatively,  we  can  also

generate  a  precision-recall  (PR)  plot  to  better  understand  the  trade-oﬀ  relationship  between  the

precision  vs  recall  for  each  group  [[80](#link_bookmark_265)–[83](#link_bookmark_286)].  While  both  of  these  plots  are  very  useful  for  visual  com-

parison  of  the  performance  trade-oﬀs  between  diﬀerent  algorithms  (or  even  diﬀerent  formulations

or  parametrizations  of  the  same  algorithm)  it  is  often  very  useful  to  be  able  to  distil  these  visual

plots  into  a  single  performance  metric  to  make  a  ﬁnal  determination  on  algorithm  performance.

This  is  especially  true  in  the  case  where  an  algorithm  designer  would  like  to  automate  the  search

for  the  optimal  algorithm  (parametrization  or  formulation).

Both the ROC and precision-recall plots are easily interpretable and can be summarized by mea-

suring  the  areas  under  each  of  the  ROC  and  precision-recall  curves  (ROC  AUC  and  precision-recall

AUC  respectively)  [[79](#link_bookmark_263)–[82](#link_bookmark_264)].  Plots  with  an  AUC  value  closer  to  1  represent  better  overall  classiﬁers

(AUC  values  lower  than  0.5  indicate  performance  that  is  worse  than  random  classiﬁcation).  AUC

values  are  convenient  for  comparing  multiple  ROC  or  precision-recall  plots  but  are  generally  not

30

recommended  as  the  sole  decision  criteria  [[79](#link_bookmark_263)–[82](#link_bookmark_264)].  This  is  simply  because  a  unidimensional  metric

1     2

*k*

1*,c*

2*,c*

*n,c*

1*,c*

1*,n*

inevitably  obfuscates  the  underlying  curve  (recall  the  moral  of  Anscombe’s  quartet)  [[88](#link_bookmark_294)].

**Binary  Classiﬁer  Performance  Evaluation**In  the  case  of  a  binary  (mutually  exclusive  two-

class)  classiﬁer  (e.g.   where  the  decision  must  be  made  to  classify  a  patient  into  NYHA  class  II

or  III  only),  which  can  be  summarized  using  a  single  ROC  or  precision-recall  (PR)  curve  (since,

by  symmetry,  improving  the  classiﬁcation  of  one  class  reduces  the  misclassiﬁcation  of  the  second

and  vice  versa),  selecting  either  the  ROC  AUC  or  precision-recall  AUC  is  natural,  expedient  and

generally very eﬀective [[79](#link_bookmark_263)–[82](#link_bookmark_264)].  However the selection of ROC AUC over precision-recall (PR) AUC

depends  on  the  underlying  class  balance  of  the  dataset.  David  et  al.  proved  a  ’surprising  theorem

that a curve dominates in ROC space if and only if it dominates in PR space’ but ’that an algorithm

that  optimizes  the  area  under  the  ROC  curve  is  not  guaranteed  to  optimize  the  area  under  the  PR

curve’  [[81](#link_bookmark_297)].  Precision-recall  curves,  and  by  extension  the  PR  AUC,  are  more  appropriate  for  use  in

datasets with highly skewed class distributions (e.g.  a prevalent negative class with few positive class

-  very  common  in  the  ﬁeld  of  information  retrieval)  [[81](#link_bookmark_297),[82](#link_bookmark_264), [84](#link_bookmark_273)].  Precision-recall  curves  however  are

more  computationally  complicated  to  solve  although  several  packages  exist  for  the  R  programming

language  (amongst  others)  to  calculate  PR  AUC,  ROC  AUC  and  all  the  various  metrics  discussed

thus  far  [[89](#link_bookmark_301), [90](#link_bookmark_302)].

**Multi-class   Classiﬁer   Performance   Evaluation**In  the  case  of  multi-class  classiﬁers  (e.g.

where  the  decision  presented  is  to  classify  a  patient  into  any  of  the  four  classes,  or  even  just

three  of  the  four  classes:   e.g.    II-IV)  the  situation  is  more  complicated.    Multi-class  classiﬁers

cannot  be  summarized  using  a  single  ROC  or  precision-recall  curve  in  the  same  way  as  binary

classiﬁers [[82](#link_bookmark_264),[84](#link_bookmark_273),[87](#link_bookmark_287),[91](#link_bookmark_306)–[93](#link_bookmark_307)].  There have been attempts by researchers to develop an equivalent AUC

metric  by  expressing  it  in  an  ’equivalent  probabilistic  form’  [[91](#link_bookmark_306)],  or  by  plotting  multi-dimensional

ROC  or  precision-recall  surfaces  -  although  this  metric  is  more  accurately  termed  the  volume-

under-the-curve  (VUC)  for  the  3  class  case,  hyper-volume-under-the-curve  (hVUC)  for  the  4  class

case,  and  so  on  [[87](#link_bookmark_287)].  None  of  these  metrics  however  has  gained  widespread  acceptance.  The  most

accepted  approach  however  appears  to  be  to  completely  forgo  calculation  of  the  AUC  in  the  ﬁrst

place  and  instead  perform  a  micro-  or  macro-average  of  an  underlying  performance  measure  (such

speciﬁcity)  [[82](#link_bookmark_264), [84](#link_bookmark_273), [92](#link_bookmark_312)].

**Macro-  and  Micro-Averaging**The  macro-average  of  a  measure,  is  calculated  in  the  same  way

as  one  might  compute  the  ’classical’  mean  of  a  the  measure  by  simply  taking  the  mean  of  the

measure  output  values.   The  micro-average  however  involves  instead  ﬁrst  taking  the  mean  of  the

inputs  to  the  measure  computation  function  and  using  those  input  mean  values  to  calculate  a  single

measure  output  value  [[86](#link_bookmark_267)].  The  diﬀerence  is  sometimes  clearer  and  more  obvious  when  expressed

mathematically.

Given a set of *k*classes *C*= *c  , c  , ...c*and a binary performance measure *B*(*m    ,m    , ...,m*)

that  can  be  calculated  for  a  given  class  *c*(and  for  each  class  *c*)  based  on  a  values  in  the  confusion

matrix  *m*to  *m*for  that  class  (which  in  our  case  would  be  the  main  confusion  metric  measures:

true positives & negatives and false positives & negatives) then the macro-average is for performance

measure  *B*is  calculated  as  [[85](#link_bookmark_314), [86](#link_bookmark_267), [94](#link_bookmark_316)]:

31

1*,j*

2*,j*

*n,j*

*k*

*k*

*B*

*macro*

=

1

*k*

*k*  
X

*j*=1

*B*(*m    ,m    , ...,m*)

(1)

In  comparison  the  micro-average  is  calculated  as  [[85](#link_bookmark_314), [86](#link_bookmark_267), [94](#link_bookmark_316)]:

*k*  
 X          X              X

*Bmicro*= *B*(      *m*1*,j ,       m*2*,j , ...,       mn,j*)

(2)

*j*=1

*j*=1

*j*=1

**Eﬀects  of  Class  Imbalance**Unfortunately,  it  is  still  an  open  question  as  to  whether  micro-

averaging  or  macro-averaging  produces  more  reliable  results.  Van  Asch  quoting  a  paper  by  Sebas-

tiani  outlines  the  state  of  the  debate  in  the  early  2000’s  [[86](#link_bookmark_267)]:

There  is  no  complete  agreement  among  authors  on  which  is  better.  Some  believe  that

micro-averaged  performance  is  somewhat  misleading  (...)  because  more  frequent  topics

are weighted heavier in the average[Wiener et al.  1995, page 327] and thus favour macro-

averaging.    Others  (actually,  the  majority  of  researchers)  believe  that  topics  should

indeed  count  proportionally  to  their  frequence,  and  thus  lean  towards  micro-averaging.

[[95](#link_bookmark_321)]

The  choice  is  sometimes  performance  measure  dependent;  Forman  et  al.,  based  on  series  of  simula-

tions  they  performed,  advanced  the  majority  position  that  micro-averaging  should  be  the  preferred

method  for  calculating  the  F  score  (a  common  alternative  to  accuracy  in  the  information  retrieval

domain) since it is more unbiased in the common case where class imbalance exist in those datasets,

especially  where  that  class  imbalance  is  large  or  the  classiﬁer  is  inaccurate  in  the  ﬁrst  place  [[85](#link_bookmark_314)].

A  decade  and  a  half  later,  the  issue,  at  least  in  practice,  appears  to  still  not  be  settled.   A  quick

review  of  recent  literature  shows  that  researchers  appear  to  prefer  to  simply  use  and  report  both

the  micro-  and  macro-averaged  results  for  their  metrics  (usually  precision,  recall,  accuracy  and  F

score)  [[96](#link_bookmark_323)–[99](#link_bookmark_324)].   Some  have  even  eschewed  micro-  and  macro-averaging  in  favor  of  developing  new

metrics,  such  as  the  Multiclass  Performance  Score  (MPS),  designed  speciﬁcally  for  use  with  multi

class  classiﬁers  [[93](#link_bookmark_307)].  In  keeping  with  common  practice  we  plan  to  report  on  the  micro-  and  macro-

averaged  precision,  recall,  accuracy  and  F  score  of  our  algorithm.  But,  in  our  case,  given  our  intent

to  integrate  our  classiﬁcation  algorithm  into  a  larger  clinical  decision  support  application  at  the

Heart Failure Clinic we do not have the freedom of waﬄing between all of these measures but rather

must  make  a  determination  of  which  single  measure  is  most  appropriate  for  the  evaluation  of  our

candidate  algorithms  for  optimization.  To  do  this  we  return  to  the  basics.

The  fundamental  diﬀerence  between  macro-  and  micro-averaging  is  how  the  weighting  is  dis-

tributed.  In  macro-averaging  equal  weight  is  given  to  each  class.  In  micro-averaging  equal  weight

is  given  to  each  individual  patient  classiﬁcation  decision.   To  Forman  et  al.’s  point,  it’s  not  that

micro-averaging  is  more  unbiased  so  much  as  it’s  particular  bias  was  more  suitable  given  the  ﬁeld  of

application  considered  by  Forman  et  al.  (namely  document  retrieval  where  giving  adequate  weight-

ing  to  less  frequently  occurring  individual  results  -  typically  the  documents  to  be  retrieved  -  is  more

preferable).  In  a  sense,  every  measure  is  biased  -  the  question  is  simply  which  bias  is  more  helpful

for  the  application  in  question.

In  our  case  we  also  expect  to  have  an  imbalance  of  persons  in  each  class  for  various  reasons:

a)  sicker  patients  (closer  to  class  IV)  are  more  likely  to  die  and  therefore  those  classes  are  likely

to  be  more  under  represented,

32

b)  the  UHN  Heart  Failure  clinic  typically  sees  more  advanced  cases  of  Heart  Failure  and  is  less

[1](#link_bookmark_326)

likely  to  see  patients  in  the  lowest  class  (class  I)  making  this  class  more  likely  to  be  under

represented  ,

c)  based  on  our  previous  experience  with  the  Medly  tele-monitoring  platform,  class  IV  patients

are  less  likely  to  be  prescribed  the  platform  as  a  clinical  intervention  (and  thus  be  eligible  for

the  study)  since  physicians  perceive  them  to  be  less  likely  to  beneﬁt  from  the  intervention.

As  a  result,  this  class  is  again  more  likely  to  be  under  represented.

Of  course  we  will  need  to  examine  the  ﬁnal  recruitment  cohort  to  conﬁrm  that  our  expectations

do  indeed  turn  out.    Have  established  that  “microaveraged  results  are  ...    really  a  measure  of

eﬀectiveness  on  large  classes  in  a  test  collection.   To  get  a  sense  of  eﬀectiveness  on  small  classes,

[one]  should  compute  macroaveraged  results”  [[100](#link_bookmark_327)].   This  would  indicate  that  the  use  of  macro-

averaging  will  provide  the  most  suitable  bias  for  our  particular  situation  since,  as  discussed,  NYHA

class  IV  is  likely  to  be  under  represented  in  our  population.  Since  macro-averaging  will  treat  every

class  considered  with  equal  weighting  optimizing  our  algorithm  using  a  macro-averaged  metric  is

more likely to ensure that classiﬁcation performance of a smaller NYHA class IV will be treated with

the  same  priority  as  any  other  class.  Optimizing  using  micro-averaging  would  instead  preferentially

weigh  improvements  (or  decreases)  in  algorithm  performance  that  largely  aﬀect  those  classes  whose

patients are more represented in the dataset (NYHA class II or III) which we suggest is undesirable.

Of course  we do  note that the use  of microsimulation  allows us  to compensate for any signiﬁcant

imbalances in our development dataset by allowing us to simulate (i.e.  generate) a balanced dataset

should  we  so  choose.  The  issue  of  micro-  vs  macro-simulation  therefore  is  perhaps  less  critical  for

this  particular  study  as  it  may  be  for  future  (external)  validation  studies.

**Misclassiﬁcation  Costs**So putting aside the issue of micro- and macro-averaging for a moment.

Thus  far  we  have  entirely  avoided  the  issue  of  misclassiﬁcation  cost.  For  binary  classiﬁcation  prob-

lems  misclassiﬁcation  cost  is  generally  glossed  over  since  there  is  only  a  single  cost  for  incorrect

classiﬁcation.   This  cost  can  be  taken  into  account  after  computation  of  performance  measures.

However,  in  certain  multi  class  classiﬁcation  problems  (such  as  this  one)  certain  types  of  misclas-

siﬁcation  can  been  deemed  to  be  more  costly  than  others  and  so  there  is  not  necessarily  a  single

cost  for  misclassiﬁcation.  In  our  case:  misclassiﬁcation  of  a  class  II  patient  into  class  IV  is  a  more

severe  error,  and  thereby  most  costly  error,  than  misclassiﬁcation  of  a  class  II  patient  into  class

III.  It  is  a  relatively  easy  task  to  design  an  algorithm  to  account  for  this  and  so  the  diﬃculty  here

mostly revolves around how to quantify the exact costs of diﬀerent degrees of misclassiﬁcation (costs

which  also  could  very  well  change  over  time)  [[93](#link_bookmark_307)].   This  challenge  is  rendered  even  more  diﬃcult

since  classiﬁcation  of  patients  into  NYHA  class,  as  already  mentioned,  is  already  an  unreliable

art  [[43](#link_bookmark_29), [44](#link_bookmark_141), [46](#link_bookmark_142), [47](#link_bookmark_30)].   And  where  there  is  already  a  relative  paucity  of  published  research  regarding

the  reliability  of  NYHA  classiﬁcation  by  physicians  there  is  complete  drought  when  it  comes  to

attempts  to  quantify  the  real  costs  of  misclassiﬁcation.  Fortunately  selecting  an  incorrect  misclas-

siﬁcation  cost  in  this  context  will  not  actually  reduce  the  accuracy  of  a  selected  output  algorithm.

Instead  it  would  direct  the  optimization  algorithm  to  search  for  a  suboptimal  algorithm,  not  that

any optimization algorithm even guarantees that it has found the single most (i.e.  globally) optimal

algorithm  in  the  ﬁrst  place  [[101](#link_bookmark_333)].  Given  the  aforementioned  consideration  and  the  lack  of  readily

available data documenting concrete misclassiﬁcation costs we propose the following as a reasonable

1

which  again,  is  also  why  although  the  general  approach  proposed  is  easily  extensible  to  this  class,  NYHA  class  I

patients  are  generally  disregarded  in  our  study

33

selection  of  costs  to  use  as  a  starting  point  in  this  study.  Of  course,  we  invite  the  interested  reader

*c*

*c*

*c*

*c*

*c*

   


   


*c*

*c*

*j*

1

2

3

*c*

*c*

*c*

*micro*

*n*

*c*

.

.

.

*c*

*macro*

*n*

.

.

.

*n*

*c*

*c*

*F*

*r*

*c*

Algorithm

Classiﬁcation

to  improve  on  our  approach.

**Derivation  of  a  Performance  Metric  that  accounts  for  misclassiﬁcation  costs  and  class**

**imbalance**Recall  that  *true  positive  rate*(*sensitivity*/*recall*)  measures  the  proportion  of  patients

correctly  identiﬁed  as  belonging  to  a  particular  test  class.   Given  a  confusion  matrix  of  *n*classes

which  can  be  encoded  as  a  *n   ×−n*mathematical  matrix  *M*as  follows,  demonstrated,  without  loss

of  generality,  using  a  *n*= 3  class  confusion  matrix:

Clinical  Classiﬁcation

II

III

IV





II

*a*

*b*

*c*

*a    b     c*

III

*d*

*e*

*f*

=*⇒−Mc*= *d    e    f*





IV

*g*

*h*

*i*

*g    h*

*i*

Let  the  grand  sum  of  the  elements  of  *M*be  equal  to  *k*,  and  the  sum  of  the  elements  in  each  column

*j*of  *M*be  equal  to  *k*.  For  our  example:

*k*= *a*+ *b*+ *c*

*k*= *d*+ *e*+ *f*

*k*= *g*+ *h*+ *i*

The  rectangular  *n   ×−n*averaging  matrix  *K*for  *M*must  be  deﬁned  either  as  :





1    1

*. . .*1

1

1

   


1    1

*. . .*

   
 

1

*K*= *K*

=

*k*

*J*=

*k*

   


   


.   
.

.

.

.

.

.

.   
.

   


   


,  for  micro-averaging

1    1

*. . .*1

*n  ×n*

*c*

or





1

*k*

1

0

*. . .*

0

1

   




0

1

*. . .*

0

   




*k*

2

*K*= *K*

=

*n*

*c*

*J*

*c*





   


.

.

.   
.

.

.

.

.   
.





   


,  for  macro-averaging

0

0

*. . .*

*k*

1

*c       n  ×n*

*c*

We  can  therefore  represent  the  calculation  of  the  *true  positive  rate*or  *recall*of  *M*as  Frobenius

inner  product  (*h., .i−*;  which  returns  the  sum  of  the  component-wise  inner  product  of  two  matrices)

of  a  weighting  matrix  *W*,  and  the  Hadamard  product  (i.e.  element-wise  product:  )  of  the  desired

micro-  or  macro-averaging  matrix  *K*and  the  confusion  matrix  *M*:

34

*recall*= *hW  ,K     M  i−*

*r*

*c   F*

*r*

.

.

.

*c*

*c*

*r*

*c   F*

*r*

*micro*

*c   F*

*r*

*c   F*

*r*

*c*

Algorithm

Classiﬁcation

(3)

The  weight  matrix  for  the  (classic  deﬁnition  of  the)  *true  positive  rate*would  be  as  follows:





1    0

*. . .*0

   


0    1

*. . .*

   
 

0

*W*=

   


   


.

.

.   
.

.

.

.

.   
.

   


   


0    0

*. . .*1

*n  ×n*

*c*

It  may  be  clearer  to  see  why  this  is  the  case  given  an  example.   The  above  implies  the  following

weighting  for  our  3  class  example  (displayed  in  a  confusion  matrix  grid).  Moving  forward  we  will

use  this  confusion  matrix  inspired  grid  format  to  display  the  weights  since  it  makes  explicit  the

corresponding  classes  and  classiﬁcation  source  for  each  weight  value.

**Weight  Matrix  for  Classically  Deﬁned  Recall**

**Weights**Clinical  Classiﬁcation

II

III

IV

II

1

0

0

III

0

1

0

IV

0

0

1

**Demonstration  of  Derivation  of  Classically  Deﬁned  Recall**Each  true  positive  (along  the

diagonal of the table) for the class are each counted as 1 correct classiﬁcation and each false negatives

is  counted  as  contributing  0  correct  classiﬁcations  to  the  overall  *true  positive  rate*’score’.   So  for

our  example  3  class  *M*matrix  the  recall  is  calculated  as  follow.   We  use  micro-averaging  for  the

sake  of  clarity,  the  technique  is  equivalent  for  macro-averaging  but  more  clumsy  as  a  result  of  the

diﬀerent  weightings  for  each  column.  Recall  the  formula  for  recall  is:

*recall*= *hW  ,K     M  i−*

and  so  for  micro-averaging  becomes:

*recall*= *hW  ,K            M  i−*

substituting  in  for  *K*

*macro*

,

*recall*= *hW  ,*

1

*k*

*M  i−*

substituting  in  both  *W*and  *M*,

35



   
 

*h**w*

*i−*

*x*) =

*a,*1

*b,*2

*d,*1

*e,*2

*f,*3

*g,*1

*h,*2

   

      
   

*F*

*a,*1

*b,*2

*c,*3

*e,*1

*e,*2

*f,*3

*g,*1

*h,*2

*i,*3

*a,*1

*e,*2

*i,*3

*d,*1

*g,*1

*b,*2

*h,*2

*c,*3

*f,*3

*tp*

*a,*1

*e,*2

*i,*3

*fn*

*d,*1

*g,*1

*b,*2

*h,*2

*c,*3

*f,*3

*tp*

*fn*

*tp*

*fn*

*class*1

*class*2

*class*3

*tp*

*fn*

*tp*

*fn*

*classic*

*c     c*

   



*w*

*w        w*

*c,*3

*a    b     c*

*recall*=

1

*k*

   

*w        w       ,*

   
 

*d    e    f*



*w        w*

*w*

*i,*3

*g    h*

*i*

then  the  Frobenius  inner  product  of  the  matrices  evaluates  to:

*recall*=

1

*k*

(*aw*+ *bw*+ *cw*+ *dw*+ *ew*+ *fw*+ *gw*+ *hw*+ *iw*)

or  equivalently:

*recall*=

1

*k*

(*aw*+ *ew*+ *iw*) +

1

*k*

((*dw*+ *gw*) + (*bw*+ *hw*) + (*cw*+ *fw*))     (4)

We  pause  to  take  note  that  if  we  let:

*w*= *w*= *w*= *w*

and:

*w*= *w*= *w*= *w*= *w*= *w*= *w*

and  recalling  that  *k*is  simply  the  sum  of  all  the  cells:

*k*= *a*+ *b*+ *c*+ *d*+ *e*+ *f*+ *g*+ *h*+ *i*

then  equation  [4 b](#link_bookmark_334)

*recall*=

*w*(*a*+ *e*+ *i*) + *w*((*d*+ *g*) + (*b*+ *h*) + (*c*+ *f*))

*a*+ *b*+ *c*+ *d*+ *e*+ *f*+ *g*+ *h*+ *i*

(5)

i.e.:

*recall*=

*w*(*a*+ *e*+ *i*) + *w*((*d*+ *g*) + (*b*+ *h*) + (*c*+ *f*))

(*a*+ *e*+ *i*) + ((*d*+ *g*) + (*b*+ *h*) + (*c*+ *f*))

(6)

Here  we  have  essentially  grouped  the  true  positive  counts:

*True  Positives*= *TP*= *a*+ *e*+ *i*

and  false  negative  counts  (sub-grouped  for  each  class):

*False  Negatives*= *FN*= *FN*

+ *FN*

+ *FN*

= (*d*+ *g*) + (*b*+ *h*) + (*c*+ *f*)

i.e.

*recall*=

*w   TP*+ *w    FN*

*TP*+ *FN*

(7)

and  for  *w*=  1  and  *w*=  0  we  can  clearly  see  here  that  we  have  simply  replicated  the  classic

*True  Positive  Rate*,  *recall*formula:

*recall|−*

= *P*( ˆ *|x*

*TP*

*TP*+ *FN*

(8)

36

**Weight  Matrix  for  Modiﬁed  Recall**But,  we  have  gained  added  ﬂexibility  because  we  now

*a,*1

*b,*2

*c,*3

*tp*

*fn*

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

have  weight  parameters  (*w*,  *w*,  *w*and  so  on)  that  we  can  now  manipulate  to  alter  how  the

recall  function  scores  each  individual  type  of  classiﬁcation  or  misclassiﬁcation  instead  of  a  blanket

*w*=  1  for  each  correct  classiﬁcation  (which  is  not  inherently  problematic  in  our  situation)  and

*w*= 0  for  each  incorrect  classiﬁcation  (which  is  the  problem  we  are  attempting  to  resolve).

In order to penalize more egregious misclassiﬁcations we propose the use of the following adjusted

weight  matrix  (for  a  3  class  classiﬁer.  A  4  class  classiﬁer  would  require  a  diﬀerent  weight  matrix):

**Weights**Clinical  Classiﬁcation

II

III

IV

II

+1

0

*−*1

III

0

+1

0

IV

*−*1

0

+1

**Demonstration  of  the  eﬀects  of  the  use  of  Modiﬁed  Recall  on  algorithm  evaluation**

The  above  weight  matrix  is  derived  from  the  following  rules,  which  we  simultaneous  demonstrate

using  an  example  sub-set  of  the  a  confusion  matrix  sub-set  (shown  immediately  below)  and  it’s

corresponding  *True  Positive  Rate*score  -  we  let  the  default  score  equal  *x*:

Clinical  Classiﬁcation

II

III

IV

II

*a*

*...*

*...*

+1 *∗−a*

III

*d*

*...*

*...*

=*⇒−*

+0 *∗−d*

IV

*g*

*...*

*...*

*−*1 *∗−g*

score:  1*a −−*1*g*= *x*

1.  In  keeping  with  the  classic  *True  Positive  Rate*formula  an  increase  in  the  number  of  misclas-

siﬁcations,  *ceteris  paribus*,  should  not  cause  an  increase  in  the  *True  Positive  Rate*score.

increase  in  the  number  of  class  II  patient’s  misclassiﬁed  as  class  III:

Clinical  Classiﬁcation

II

III

IV

II

*a*

*...*

*...*

+1 *∗−a*

III

*d*+ *n*

*...*

*...*

=*⇒−*+0 *∗−*(*d*+ *n*)

IV

*g*

*...*

*...*

*−*1 *∗−g*

score:  *x*+ 0

increase  in  the  number  of  class  II  patient’s  misclassiﬁed  as  class  IV:

37

Clinical  Classiﬁcation

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

II

III

IV

II

*a*

*...*

*...*

+1 *∗−a*

III

*d*

*...*

*...*

=*⇒−*

+0 *∗−d*

IV

*g*+ *n*

*...*

*...*

*−*1 *∗−*(*g*+ *n*)

score:  *x −−n*

2.  A  decrease  in  misclassiﬁcation  severity  for  a  given  patient,  *ceteris  paribus*,  should  cause  an

increase  in  the  *True  Positive  Rate*.

class  II  patient  misclassiﬁcation  improvement  to  III  from  IV:

Clinical  Classiﬁcation

II

III

IV

II

*a*

*...*

*...*

+1 *∗−a*

III

*d*+ 1

*...*

*...*

=*⇒−*+0 *∗−*(*d*+ 1)

IV

*g −−*1

*...*

*...*

*−*1 *∗−*(*g −−*1)

score:  *x*+ 1

class  II  patient  misclassiﬁcation  improvement  to  II  from  III:

Clinical  Classiﬁcation

II

III

IV

II

*a*+ 1

*...*

*...*

+1 *∗−*(*a*+ 1)

III

*d −−*1

*...*

*...*

=*⇒−*+0 *∗−*(*d −−*1)

IV

*g*

*...*

*...*

*−*1 *∗−g*

score:  *x*+ 1

3.  An  increase  in  misclassiﬁcation  severity  for  a  given  patient,  *ceteris  paribus*,  should  cause  a

decrease  in  the  *True  Positive  Rate*.

class  II  patient  misclassiﬁcation  deterioration  to  III  from  II:

38

Clinical  Classiﬁcation

[2](#link_bookmark_335)

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

II

III

IV

II

*a −−*1

*...*

*...*

+1 *∗−*(*a −−*1)

III

*d*+ 1

*...*

*...*

=*⇒−*+0 *∗−*(*d*+ 1)

IV

*g*

*...*

*...*

*−*1 *∗−g*

score:  *x −−*1

class  II  patient  misclassiﬁcation  deterioration  to  IV  from  III:

Clinical  Classiﬁcation

II

III

IV

II

*a*

*...*

*...*

+1 *∗−a*

III

*d −−*1

*...*

*...*

=*⇒−*+0 *∗−*(*d −−*1)

IV

*g*+ 1

*...*

*...*

*−*1 *∗−*(*g*+ 1)

score:  *x −−*1

4.  The  health  of  any  arbitrary  patient  should  be  considered  as  having  at  least  the  same  value  as

the  health  of  any  arbitrary  other  patient  .  In  other  words,  *Ceteris  paribus*,  an  algorithm  that

is  able  to  decrease  the  misclassiﬁcation  severity  from  baseline  for  a  number  of  patients,  at

the  cost  of  increasing  the  misclassiﬁcation  severity  from  baseline  for  an  equivalent  number  of

other  patients  should  be  considered  strictly  as  not  performing  any  better  than  an  equivalent

algorithm  that  simply  classiﬁes  patients  at  those  baselines.

class  II  patient  misclassiﬁcation  improvement  to  III  from  IV  and

class  II  patient

misclassiﬁcation  deterioration  to  IV  from  III:

Clinical  Classiﬁcation

II

III

IV

II

*a*

*...*

*...*

+1 *∗−a*

III

*d*+ 1 *−−*1

*...*

*...*

=*⇒−*+0 *∗−*(*d*+ 1 *−−*1)

IV

*g −−*1 + 1

*...*

*...*

*−*1 *∗−*(*g −−*1 + 1)

score:  *x*+ 0

2

This  sort  of  accounting  of  human  life  is  admittedly  rather  crude,  having  a  tendency  to  lead  to  a,  probably,  over

simplistic  and  myopic  utilitarian  view  of  justice,  equality  and  fairness  [[102](#link_bookmark_336)].   It  is  the  author’s  admission  that  this

is  not  necessarily  an  ideal  and  perhaps  a  future  reader  from  a  more  enlightened  time  will  be  able  to  improve  on

this  crude  assumption  if  ever  society  is  able  to  solve  the  complex  calculus  of  human  life  valuation.   However  this

assumption  does  at  least  reﬂect  a  not  uncommonly  accepted  viewpoint,  or  at  least  an  ideal,  of  present  (western)

cultural  and  societal  thought  and  so  is  at  least,  in  the  author’s  humble  opinion,  is  a  not  unreasonable  starting  point.

39

class  II  patient  misclassiﬁcation  improvement  to  II  from  III  and

[3](#link_bookmark_337)

Algorithm

Classiﬁcation

Algorithm

Classiﬁcation

class  II  patient

misclassiﬁcation  deterioration  to  III  from  II:

Clinical  Classiﬁcation

II

III

IV

II

*a*+ 1 *−−*1

*...*

*...*

+1 *∗−*(*a*+ 1 *−−*1)

III

*d −−*1 + 1

*...*

*...*

=*⇒−*+0 *∗−*(*d −−*1 + 1)

IV

*g*

*...*

*...*

*−*1 *∗−g*

score:  *x*+ 0

class  II  patient  misclassiﬁcation  improvement  to  II  from  III  and

class  II  patient

misclassiﬁcation  deterioration  to  IV  from  III:

Clinical  Classiﬁcation

II

III

IV

II

*a*+ 1

*...*

*...*

+1 *∗−*(*a*+ 1)

III

*d −−*2

*...*

*...*

=*⇒−*+0 *∗−*(*d −−*2)

IV

*g*+ 1

*...*

*...*

*−*1 *∗−*(*g*+ 1)

score:  *x*+ 0

And  so  using  the  adjusted  weight  matrix  instead  of  an  identify  matrix  (i.e.  the  weight  matrix

for  the  classic  deﬁnition  of  *recall*)  we  can  instead  guide  to  optimization  algorithm  to  not  only  to

prefer  algorithms  that  are  strictly  better  at  classifying  patients  into  their  corresponding  class  but

when  patients  are  misclassiﬁed  to  prefer  algorithms  that  misclassify  them  less  incorrectly.

**Summary**To  summarize:  how  speciﬁcally  we  report  and  assess  the  performance  of  our  classiﬁer

will  depend  primarily  on  how  many  classes  it  supports.   Either  way  though  the  selection  a  ﬁnal

candidate  algorithm  we  will  use  machine  optimization  (although  the  exact  method  is  not  yet  de-

termined    )  to  iterate  though  various  parametrizations  of  the  classiﬁcation  algorithm  to  identify

more  performant  versions.  For  the  purpose  of  optimization  the  performance  of  the  algorithm  will

be  evaluated  based  on  a  relevant  performance  metric  depending  on  how  many  classes  it  supports.

For  a  binary  classiﬁer  we  will  use  the  area  under  the  receiver  operating  characteristic  curve  (AUC

ROC).  For  a  multi-class  classiﬁer  we  will  use  a  modiﬁcation  to  the  *recall*performance  metric  where

more  extreme  misclassiﬁcation  of  patients  will  be  more  heavily  penalized  instead  of  penalized  iden-

tically  to  less  extreme  misclassiﬁcations  (correct  classiﬁcations  will  be  treated  identically  to  the

classical  interpretation  of  the  performance  measure).   This  modiﬁcation  will  allow  us  to  select  for

algorithms that are not just more correct, but also ’less wrong’.  We will report standard measures of

predictive  accuracy  for  the  ﬁnal  classiﬁcation  algorithm  developed  whether  binary  (class  II  or  III),

or  multi-class  (II,  III  and  IV)  -  dependant  on  our  ability  to  capture  class  IV  patients  for  analysis.

3

Possible  candidates  include:  simulated  annealing,  genetic  algorithms,  amongst  others  [[101](#link_bookmark_333)]

40

These  measures  will  include  the  (classically  deﬁned)  *sensitivity/recall*,  *speciﬁcity*,  *fall-out*,  *precision*

and  *F-score*.  These  standard  measures  will  be  reported  as  raw  measures  for  a  binary  classiﬁer  and

as  macro-  and  micro-averaged  measures  for  a  multi-class  classiﬁer.  We  note  that  the  assessment  of

the  ﬁnal  classiﬁcation  algorithm  will  be  performed  by  running  the  classiﬁer  against  the  (internal)

validation  dataset  (which  is  the  same  as  the  development  dataset)  and  as  a  result  we  expect  that

evaluation  of  algorithm  performance  to  be  optimistic.

**2.8     Development  vs.  Validation**

Since  we  reuse  a  single  dataset  the  setting,  eligibility  criteria,  outcome,  and  predictors  remain

unchanged  for  both  datasets.   Since  model  development  and  performance  assessment  is  also  per-

formed  using  the  same  dataset  we  expect  the  reported  performance  to  be  optimistic.

**3     Results**

**3.1     Participants**

**3.2     Model  Development**

**3.3     Model  Speciﬁcation**

**3.4     Model  Performance**

**3.5     Model-updating**

**4     Discussion**

**4.1     Limitations**

**4.2     Interpretation**

**4.3     Implications**

**5     Other  Information**

**5.1     Supplementary  Information**

**5.2     Funding**

Funding  for  this  project  was  provided  by  the:

Centre  for  Global  eHealth  Innovation  (Heart  Function  Clinic?)

Canadian  Institutes  of  Health  Research  (Government  of  Canada)

Government  of  Ontario

University  of  Toronto

41

**References**

[1]  M.    R.    Mehra    and    J.    Butler,     “Heart    Failure:       A    Global    Pandemic    and    Not

Just    a    Disease    of    the    West,”    *Heart    Failure    Clinics*,    vol.    11,    no.    4,    pp.    xiii–

xiv,   oct   2015.   [Online].   Available:    [http://www.ncbi.nlm.nih.gov/pubmed/26462110http:](http://www.ncbi.nlm.nih.gov/pubmed/26462110%20http://linkinghub.elsevier.com/retrieve/pii/S1551713615000690)

[//linkinghub.elsevier.com/retrieve/pii/S1551713615000690](http://www.ncbi.nlm.nih.gov/pubmed/26462110%20http://linkinghub.elsevier.com/retrieve/pii/S1551713615000690)

[2]  Heart and Stroke Foundation, “2016 Report on the Health of Canadians:  The Burden of Heart

Failure,”  2016.  [Online].  Available:   [http://www.heartandstroke.com/atf/cf/%7B99452d8b-](http://www.heartandstroke.com/atf/cf/%7B99452d8b-e7f1-4bd6-a57d-b136ce6c95bf%7D/2016-HEART-REPORT.PDF)

[e7f1-4bd6-a57d-b136ce6c95bf%7D/2016-HEART-REPORT.PDF](http://www.heartandstroke.com/atf/cf/%7B99452d8b-e7f1-4bd6-a57d-b136ce6c95bf%7D/2016-HEART-REPORT.PDF)

[3]  E.  Seto,  K.  J.  Leonard,  J.  a.  Cafazzo,  C.  Masino,  J.  Barnsley,  and  H.  J.  Ross,  “Self-care

and  quality  of  life  of  heart  failure  patients  at  a  multidisciplinary  heart  function  clinic.”

*The  Journal  of  cardiovascular  nursing*,  vol.  26,  no.  5,  pp.  377–85,  2011.  [Online].  Available:

<http://www.ncbi.nlm.nih.gov/pubmed/21263339>

[4]  S.  Lawrence,  “Canada  is  failing  our  heart  failure  patients  -  Heart  and  Stroke  Foundation

of   Canada,”   2016.   [Online].   Available:    [http://www.heartandstroke.com/site/apps/nlnet/](http://www.heartandstroke.com/site/apps/nlnet/content2.aspx?c=ikIQLcMWJtE&b=3485819&ct=14816887)

[content2.aspx?c=ikIQLcMWJtE&b=3485819&ct=14816887](http://www.heartandstroke.com/site/apps/nlnet/content2.aspx?c=ikIQLcMWJtE&b=3485819&ct=14816887)

[5]  J.  Cox  and  C.  D.  Naylor,  “The  Canadian  Cardiovascular  Society  Grading  Scale  for  Angina

Pectoris:  Is It Time for Reﬁnements?”  *Annals of Internal Medicine*, vol. 117, no. 8, p. 677, oct

1992.  [Online].  Available:  <http://annals.org/article.aspx?doi=10.7326/0003-4819-117-8-677>

[6]  C.  Raphael,  C.  Briscoe,  J.  Davies,  Z.  Ian  Whinnett,  C.  Manisty,  R.  Sutton,  J.  Mayet,  and

D. P. Francis, “Limitations of the New York Heart Association functional classiﬁcation system

and self-reported walking distances in chronic heart failure,” *Heart*, vol. 93, no. 4, pp. 476–482,

apr  2007.  [Online].  Available:  <http://heart.bmj.com/cgi/doi/10.1136/hrt.2006.089656>

[7]  J.  A.  Bennett,  B.  Riegel,  V.  Bittner,  and  J.  Nichols,  “Validity  and  reliability  of  the  NYHA

classes  for  measuring  research  outcomes  in  patients  with  cardiac  disease,”  *Heart  and  Lung:*

*Journal  of  Acute  and  Critical  Care*,  vol.  31,  no.  4,  pp.  262–270,  2002.

[8]  Heart    Foundation,    “New    York    Heart    Association    (NYHA)    Classiﬁcation,”    p.    1,

2014.   [Online].   Available:     [http://www.heartonline.org.au/media/DRL/New](http://www.heartonline.org.au/media/DRL/New_York_Heart_Association_(NYHA)_classification.pdf)

[York](http://www.heartonline.org.au/media/DRL/New_York_Heart_Association_(NYHA)_classification.pdf)

[Heart](http://www.heartonline.org.au/media/DRL/New_York_Heart_Association_(NYHA)_classification.pdf)

[Association](http://www.heartonline.org.au/media/DRL/New_York_Heart_Association_(NYHA)_classification.pdf)

[(NYHA)](http://www.heartonline.org.au/media/DRL/New_York_Heart_Association_(NYHA)_classification.pdf)

[classiﬁcation.pdf](http://www.heartonline.org.au/media/DRL/New_York_Heart_Association_(NYHA)_classification.pdf)

[9]  American    Heart    Association,     “Classes    of    Heart    Failure,”     2015.    [Online].    Avail-

able:

[http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/](http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp#.WBY_myTla2s)

[Classes-of-Heart-Failure](http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp#.WBY_myTla2s)

[UCM](http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp#.WBY_myTla2s)

[306328](http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp#.WBY_myTla2s)

[Article.jsp#.WBY](http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp#.WBY_myTla2s)

[myTla2s](http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp#.WBY_myTla2s)

[10]  A.   Ahmed,   W.   S.   Aronow,   and   J.   L.   Fleg,   “Higher   New   York   Heart   Association

classes   and   increased   mortality   and   hospitalization   in   patients   with   heart   failure   and

preserved   left   ventricular   function.”   *American   heart   journal*,    vol.   151,    no.   2,    pp.

444–50,  feb  2006.  [Online].  Available:  [http://www.ncbi.nlm.nih.gov/pubmed/16442912http:](http://www.ncbi.nlm.nih.gov/pubmed/16442912%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2771182)

[//www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2771182](http://www.ncbi.nlm.nih.gov/pubmed/16442912%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2771182)

[11]  L.  Goldman,  B.  Hashimoto,  E.  F.  Cook,  and  A.  Loscalzo,  “Comparative  reproducibility

and  validity  of  systems  for  assessing  cardiovascular  functional  class:   advantages  of  a  new

speciﬁc  activity  scale,”  *Circulation*,  vol.  64,  no.  6,  pp.  1227–1234,  1981.  [Online].  Available:

<http://circ.ahajournals.org/cgi/doi/10.1161/01.CIR.64.6.1227>

42

[12]  B.   A.   Williams,   S.   Doddamani,   M.   A.   Troup,   A.   L.   Mowery,   C.   M.   Kline,   J.   A.

Gerringer,   and  R.  T.  Faillace,   “Agreement  between  heart  failure  patients  and  providers

in  assessing  New  York  Heart  Association  functional  class,”  *Heart  and  Lung  The  Journal*

*of   Acute   and   Critical   Care*,   vol.   46,   no.   4,   pp.   293–299,   jul   2017.   [Online].   Available:

[http://dx.doi.org/10.1016/j.hrtlng.2017.05.001http://www.sciencedirect.com/science/](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[article/pii/S0147956317300031?](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[rdoc=1&](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[fmt=high&](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[origin=gateway&](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[docanchor=](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[&md5=b8429449ccfc9c30159a5f9aeaa92ﬀb&dgcid=raven](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[sd](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[recommender](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[email&ccp=y](http://dx.doi.org/10.1016/j.hrtlng.2017.05.001%20http://www.sciencedirect.com/science/article/pii/S0147956317300031?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email&ccp=y)

[13]  G.    Roul,    P.    Germain,    and    P.    Bareiss,    “Does    the    6-minute    walk    test    predict

the    prognosis    in    patients    with    NYHA    class    II    or    III    chronic    heart    failure?”

*American   Heart   Journal*,   vol.   136,   no.   3,   pp.   449–457,   sep   1998.   [Online].   Available:

<http://linkinghub.elsevier.com/retrieve/pii/S0002870398702194>

[14]  G.  J.  Balady,   R.  Arena,   K.  Sietsema,   J.  Myers,   L.  Coke,   G.  F.  Fletcher,   D.  Forman,

B.  Franklin,  M.  Guazzi,  M.  Gulati,  S.  J.  Keteyian,  C.  J.  Lavie,  R.  Macko,  D.  Mancini,  and

R.  V.  Milani,  “AHA  Scientiﬁc  Statement  Clinician’s  Guide  to  Cardiopulmonary  Exercise

Testing  in  Adults  A  Scientiﬁc  Statement  From  the  American  Heart  Association,”  *American*

*Heart  Association  Exercise  Clinical  Cardiology;   Council  on  Epidemiology  and  Prevention*.

[Online].  Available:  <http://circ.ahajournals.org/content/circulationaha/122/2/191.full.pdf>

[15]  N.  Uth,  H.  Sørensen,  K.  Overgaard,  and  P.  K.  Pedersen,  “Estimation  of  VO2max  from

the   Ratio   between   HRmax   and   HRrest   -   the   Heart   Rate   Ratio   Method,”   *European*

*Journal   of   Applied   Physiology*,   vol.   91,   no.   1,   pp.   111–115,   2004.   [Online].   Available:

[http://pure.au.dk/portal/ﬁles/14557663/UTH2004.pdf](http://pure.au.dk/portal/files/14557663/UTH2004.pdf)

[16]  G.  M.  Kline,  J.  P.  Porcari,  R.  Hintermeister,  P.  S.  Freedson,  A.  Ward,  R.  F.  McCarron,

J.  Ross,  and  J.  M.  Rippe,  “Estimation  of  VO2max  from  a  one-mile  track  walk,  gender,  age,

and  body  weight.”  *Medicine  and  science  in  sports  and  exercise*,  vol.  19,  no.  3,  pp.  253–9,  jun

1987.  [Online].  Available:  <http://www.ncbi.nlm.nih.gov/pubmed/3600239>

[17]  K.  H.  Cooper,  *Aerobics*.    Bantam  Books,  1969.

[18]  S.  Saalasti  and  A.  Pulkkinen,  “Method  and  system  for  determining  the  ﬁtness  index  of  a

person,”  2012.  [Online].  Available:  <https://www.google.com/patents/US20140088444>

[19]  N.   F.   Butte,   U.   Ekelund,   and   K.   R.   Westerterp,   “Assessing   Physical   Activity   Using

Wearable   Monitors:    Measures   of   Physical   Activity,”   *Med.   Sci.   Sports   Exerc*,   vol.   44,

no.    1S,    pp.    5–12,    2012.    [Online].    Available:      [https://pdfs.semanticscholar.org/9fad/](https://pdfs.semanticscholar.org/9fad/9108623002f685aecb7ecc98f77075057384.pdf)

[9108623002f685aecb7ecc98f77075057384.pdf](https://pdfs.semanticscholar.org/9fad/9108623002f685aecb7ecc98f77075057384.pdf)

[20]  Ap507,    “Study   shows   slow   walking   pace   is   good   predictor   of   heart-related   deaths

University  of  Leicester,”  2017.  [Online].  Available:   [https://www2.le.ac.uk/news/blog/2017-](https://www2.le.ac.uk/news/blog/2017-archive/august/study-shows-slow-walking-pace-good-predictor-of-heart-related-deaths)

[archive/august/study-shows-slow-walking-pace-good-predictor-of-heart-related-deaths](https://www2.le.ac.uk/news/blog/2017-archive/august/study-shows-slow-walking-pace-good-predictor-of-heart-related-deaths)

[21]  S.  Zhao,  K.  Chen,  Y.  Su,  W.  Hua,  S.  Chen,  Z.  Liang,  W.  Xu,  Y.  Dai,  Z.  Liu,  X.  Fan,

C.  Hou,  and  S.  Zhang,  “Association  between  patient  activity  and  long-term  cardiac  death  in

patients  with  implantable  cardioverter-deﬁbrillators  and  cardiac  resynchronization  therapy

deﬁbrillators,”  *European  Journal  of  Preventive  Cardiology*,  vol.  24,  no.  7,  pp.  760–767,  2017.

[Online].  Available:  <http://journals.sagepub.com/doi/10.1177/2047487316688982>

43

[22]  R.  Abdulmajeed,  “The  Use  of  Continuous  Monitoring  of  Heart  Rate  as  a  Prognosticator  of

Readmission  in  Heart  Failure  Patients,”  Ph.D.  dissertation,  University  of  Toronto,  2016.

[23]  Z.   J.   Eapen,   M.   P.   Turakhia,   M.   V.   McConnell,   G.   Graham,   P.   Dunn,   C.   Tiner,

C.   Rich,    R.   A.   Harrington,    E.   D.   Peterson,    and   P.   Wayte,    “Deﬁning   a   Mobile

Health   Roadmap   for   Cardiovascular   Health   and   Disease,”    *Journal   of   the   American*

*Heart   Association*,    vol.   5,    no.   7,    p.   e003119,    jul   2016.   [Online].   Available:     [http:](http://jaha.ahajournals.org/lookup/doi/10.1161/JAHA.115.003119)

[//jaha.ahajournals.org/lookup/doi/10.1161/JAHA.115.003119](http://jaha.ahajournals.org/lookup/doi/10.1161/JAHA.115.003119)

[24]  D.  Wen,  X.  Zhang,  X.  Liu,  and  J.  Lei,  “Evaluating  the  Consistency  of  Current  Mainstream

Wearable  Devices  in  Health  Monitoring:   A  Comparison  Under  Free-Living  Conditions.”

*Journal  of  medical  Internet  research*,  vol.  19,  no.  3,  p.  e68,  mar  2017.  [Online].  Available:

[http://www.jmir.org/2017/3/e68/http://www.ncbi.nlm.nih.gov/pubmed/28270382](http://www.jmir.org/2017/3/e68/%20http://www.ncbi.nlm.nih.gov/pubmed/28270382)

[25]  F.  El-Amrawy,  M.  I.  Nounou,  K.  Volpp,  M.  Patel,  N.  Lin,  and  R.  Lewis,  “Are  Currently

Available  Wearable  Devices  for  Activity  Tracking  and  Heart  Rate  Monitoring  Accurate,

Precise,  and  Medically  Beneﬁcial?”   *Healthcare  Informatics  Research*,  vol.  21,  no.  4,  p.  315,

2015.  [Online].  Available:  [https://synapse.koreamed.org/DOIx.php?id=10.4258/hir.2015.21.](https://synapse.koreamed.org/DOIx.php?id=10.4258/hir.2015.21.4.315)

[4.315](https://synapse.koreamed.org/DOIx.php?id=10.4258/hir.2015.21.4.315)

[26]  H.-S.    An,    G.    C.    Jones,    S.-K.    Kang,    G.    J.    Welk,    and    J.-M.    Lee,    “How    valid

are    wearable    physical    activity    trackers    for    measuring    steps?”

*European    Journal*

*of    Sport    Science*,    vol.    17,    no.    3,    pp.    360–368,    mar    2017.    [Online].    Available:

<https://www.tandfonline.com/doi/full/10.1080/17461391.2016.1255261>

[27]  S.  E.  Bromberg,  “Consumer  Wristband  Activity  Monitors  as  a  Simple  and  Inexpensive  Tool

for  Remote  Heart  Failure  Monitoring,”  2015.

[28]  A.    Abeles,    R.    M.    Kwasnicki,    C.    Pettengell,    J.    Murphy,    and    A.    Darzi,    “The

relationship   between   physical   activity   and   post-operative   length   of   hospital   stay:     A

systematic   review,”    *International   Journal   of   Surgery*,    jul   2017.   [Online].   Available:

<http://linkinghub.elsevier.com/retrieve/pii/S1743919117305721>

[29]  D.  B.  Bornstein,  M.  W.  Beets,  W.  Byun,  G.  Welk,  M.  Bottai,  M.  Dowda,  and  R.  Pate,

“Equating  accelerometer  estimates  of  moderate-to-vigorous  physical  activity:   In  search  of

the  Rosetta  Stone,”  *Journal  of  Science  and  Medicine  in  Sport*,  vol.  14,  no.  5,  pp.  404–410,  sep

2011.  [Online].  Available:  <http://linkinghub.elsevier.com/retrieve/pii/S1440244011000752>

[30]  M.  Awais,  S.  Mellone,  and  L.  Chiari,  “Physical  activity  classiﬁcation  meets  daily  life:  Review

on  existing  methodologies  and  open  challenges,”  *Proceedings  of  the  Annual  International*

*Conference  of  the  IEEE  Engineering  in  Medicine  and  Biology  Society,  EMBS*,  vol.  2015-

Novem,  pp.  5050–5053,  2015.

[31]  M.   Jehn,    S.   Prescher,    K.   Koehler,    S.   Von   Haehling,    S.   Winkler,    O.   Deckwart,

M.   Honold,    U.   Sechtem,    G.   Baumann,    M.   Halle,    S.   D.   Anker,    and   F.   Koehler,

“Tele-accelerometry   as   a   novel   technique   for   assessing   functional   status   in   patients

with   heart   failure:     Feasibility,   reliability   and   patient   safety,”   *International   Journal*

*of    Cardiology*,    vol.    168,    pp.    4723–4728,    2013.    [Online].    Available:      [http://ac.els-](http://ac.els-cdn.com/S016752731301396X/1-s2.0-S016752731301396X-main.pdf?_tid=6725a990-9263-11e7-a20a-00000aacb35f&acdnat=1504634309_515452e6ea9b9329b9241c3c24cf211e)

[cdn.com/S016752731301396X/1-s2.0-S016752731301396X-main.pdf?](http://ac.els-cdn.com/S016752731301396X/1-s2.0-S016752731301396X-main.pdf?_tid=6725a990-9263-11e7-a20a-00000aacb35f&acdnat=1504634309_515452e6ea9b9329b9241c3c24cf211e)

[tid=6725a990-9263-](http://ac.els-cdn.com/S016752731301396X/1-s2.0-S016752731301396X-main.pdf?_tid=6725a990-9263-11e7-a20a-00000aacb35f&acdnat=1504634309_515452e6ea9b9329b9241c3c24cf211e)

[11e7-a20a-00000aacb35f&acdnat=1504634309](http://ac.els-cdn.com/S016752731301396X/1-s2.0-S016752731301396X-main.pdf?_tid=6725a990-9263-11e7-a20a-00000aacb35f&acdnat=1504634309_515452e6ea9b9329b9241c3c24cf211e)

[515452e6ea9b9329b9241c3c24cf211e](http://ac.els-cdn.com/S016752731301396X/1-s2.0-S016752731301396X-main.pdf?_tid=6725a990-9263-11e7-a20a-00000aacb35f&acdnat=1504634309_515452e6ea9b9329b9241c3c24cf211e)

44

[32]  C. Demers, R. S. McKelvie, A. Negassa, and S. Yusuf, “Reliability, validity, and responsiveness

of  the  six-minute  walk  test  in  patients  with  heart  failure,”  *American  Heart  Journal*,  vol.  142,

no.  4,  pp.  698–703,  2001.

[33]  J.  K.  Moon  and  N.  F.  Butte,  “Combined  heart  rate  and  activity  improve  estimates  of  oxygen

consumption  and  carbon  dioxide  production  rates.”  *Journal  of  applied  physiology  (Bethesda,*

*Md.  :   1985)*,  vol.  81,  no.  4,  pp.  1754–61,  oct  1996.  [Online].  Available:   [http://www.ncbi.](http://www.ncbi.nlm.nih.gov/pubmed/8904596%20http://jap.physiology.org/content/jap/81/4/1754.full.pdf)

[nlm.nih.gov/pubmed/8904596http://jap.physiology.org/content/jap/81/4/1754.full.pdf](http://www.ncbi.nlm.nih.gov/pubmed/8904596%20http://jap.physiology.org/content/jap/81/4/1754.full.pdf)

[34]  K.  Imai,  H.  Sato,  M.  Hori,  H.  Kusuoka,  H.  Ozaki,  H.  Yokoyama,  H.  Takeda,  M.  Inoue,  and

T. Kamada,  “Vagally mediated heart rate recovery after exercise is accelerated in athletes but

blunted  in  patients  with  chronic  heart  failure,”  *Journal  of  the  American  College  of  Cardiol-*

*ogy*,    vol.   24,    no.   6,    pp.   1529–1535,    nov   1994.   [Online].   Available:     [http://ac.els-](http://ac.els-cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?_tid=ed8dfdfe-3c12-11e7-b77d-00000aab0f6c&acdnat=1495143945_dcd737ecc3b5e6a44c2319032b42abbe%20http://linkinghub.elsevier.com/retrieve/pii/0735109794901503)

[cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?](http://ac.els-cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?_tid=ed8dfdfe-3c12-11e7-b77d-00000aab0f6c&acdnat=1495143945_dcd737ecc3b5e6a44c2319032b42abbe%20http://linkinghub.elsevier.com/retrieve/pii/0735109794901503)

[tid=ed8dfdfe-3c12-](http://ac.els-cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?_tid=ed8dfdfe-3c12-11e7-b77d-00000aab0f6c&acdnat=1495143945_dcd737ecc3b5e6a44c2319032b42abbe%20http://linkinghub.elsevier.com/retrieve/pii/0735109794901503)

[11e7-b77d-00000aab0f6c&acdnat=1495143945](http://ac.els-cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?_tid=ed8dfdfe-3c12-11e7-b77d-00000aab0f6c&acdnat=1495143945_dcd737ecc3b5e6a44c2319032b42abbe%20http://linkinghub.elsevier.com/retrieve/pii/0735109794901503)

[dcd737ecc3b5e6a44c2319032b42abbehttp:](http://ac.els-cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?_tid=ed8dfdfe-3c12-11e7-b77d-00000aab0f6c&acdnat=1495143945_dcd737ecc3b5e6a44c2319032b42abbe%20http://linkinghub.elsevier.com/retrieve/pii/0735109794901503)

[//linkinghub.elsevier.com/retrieve/pii/0735109794901503](http://ac.els-cdn.com/0735109794901503/1-s2.0-0735109794901503-main.pdf?_tid=ed8dfdfe-3c12-11e7-b77d-00000aab0f6c&acdnat=1495143945_dcd737ecc3b5e6a44c2319032b42abbe%20http://linkinghub.elsevier.com/retrieve/pii/0735109794901503)

[35]  J. S. Borer, M. B¨ohm, K. Swedberg, M. Komajda, J. Rey, S. Borer, I. Ford, A. Dubost-Brama,

G.  Lerebours,  and  L.  Tavazzi,  “Heart  rate  as  a  risk  factor  in  chronic  heart  failure  (SHIFT):

the  association  between  heart  rate  and  outcomes  in  a  randomised  placebo-controlled  trial,”

[*www.thelancet.com   Lanc*,   vol.   376,   pp.   886–94,   2010.   [Online].   Available:](http://www.thelancet.com)

<http://repository.um.edu.my/10125/1/1-s2.0-S0140673610612597-main.pdf>

[36]  N.  M.  Arzeno,  M.  T.  Kearney,  D.  L.  Eckberg,  J.  Nolan,  and  C.-S.  Poon,  “Heart  rate  chaos

as  a  mortality  predictor  in  mild  to  moderate  heart  failure.”  *Conference  proceedings  :   ...*

*Annual  International  Conference  of  the  IEEE  Engineering  in  Medicine  and  Biology  Society.*

*IEEE   Engineering   in   Medicine   and   Biology   Society.   Annual   Conference*,   vol.   2007,   pp.

5051–4,  aug  2007.  [Online].  Available:   [http://ieeexplore.ieee.org/document/4353475/http:](http://ieeexplore.ieee.org/document/4353475/%20http://www.ncbi.nlm.nih.gov/pubmed/18003141)

[//www.ncbi.nlm.nih.gov/pubmed/18003141](http://ieeexplore.ieee.org/document/4353475/%20http://www.ncbi.nlm.nih.gov/pubmed/18003141)

[37]  R.  Wang,  G.  Blackburn,  M.  Desai,  D.  Phelan,  L.  Gillinov,  P.  Houghtaling,  M.  Gillinov,

C.  MA,  M.  H,  L.  RMT,  T.  DJ,  E.-A.  F,  and  P.  MS,  “Accuracy  of  Wrist-Worn  Heart  Rate

Monitors,”  *JAMA  Cardiology*,  vol.  313,  no.  6,  pp.  625–626,  oct  2016.  [Online].  Available:

<http://cardiology.jamanetwork.com/article.aspx?doi=10.1001/jamacardio.2016.3340>

[38]  T.  J.  M.  Kooiman,  M.  L.  Dontje,  S.  R.  Sprenger,  W.  P.  Krijnen,  C.  P.  van  der  Schans,

and   M.   de   Groot,   “Reliability   and   validity   of   ten   consumer   activity   trackers,”   *BMC*

*Sports   Science,   Medicine   and   Rehabilitation*,   vol.   7,   no.   1,   p.   24,   dec   2015.   [Online].

Available:    [http://bmcsportsscimedrehabil.biomedcentral.com/articles/10.1186/s13102-015-](http://bmcsportsscimedrehabil.biomedcentral.com/articles/10.1186/s13102-015-0018-5%20http://www.ncbi.nlm.nih.gov/pubmed/26464801%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4603296)

[0018-5http://www.ncbi.nlm.nih.gov/pubmed/26464801http://www.pubmedcentral.nih.gov/](http://bmcsportsscimedrehabil.biomedcentral.com/articles/10.1186/s13102-015-0018-5%20http://www.ncbi.nlm.nih.gov/pubmed/26464801%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4603296)

[articlerender.fcgi?artid=PMC4603296](http://bmcsportsscimedrehabil.biomedcentral.com/articles/10.1186/s13102-015-0018-5%20http://www.ncbi.nlm.nih.gov/pubmed/26464801%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4603296)

[39]  T.  Ferguson,  A.  Rowlands,  V,  T.  Olds,  and  C.  Maher,  “The  validity  of  consumer-level,

activity  monitors  in  healthy  adults  worn  in  free-living  conditions:   a  cross-sectional  study,”

*International  Journal  of  Behavioral  Nutrition  and  Physical  Activity*,  vol.  12,  no.  1,  p.  42,  dec

2015.  [Online].  Available:  <http://www.ijbnpa.org/content/12/1/42>

[40]  F.  Attal,  S.  Mohammed,  M.  Dedabrishvili,  F.  Chamroukhi,  L.  Oukhellou,  and  Y.  Amirat,

“Physical  Human  Activity  Recognition  Using  Wearable  Sensors.”  *Sensors  (Basel,  Switzer-*

45

*land)*,  vol.  15,  no.  12,  pp.  31 314–38,  2015.  [Online].  Available:  [http://www.pubmedcentral.](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4721778&tool=pmcentrez&rendertype=abstract)

[nih.gov/articlerender.fcgi?artid=4721778&tool=pmcentrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4721778&tool=pmcentrez&rendertype=abstract)

[41]  K.    R.    Evenson,    M.    M.    Goto,    and    R.    D.    Furberg,    “Systematic    review    of    the

validity  and  reliability  of  consumer-wearable  activity  trackers,”   *International  Journal  of*

*Behavioral   Nutrition   and   Physical   Activity*,   vol.   12,   no.   1,   p.   159,   dec   2015.   [Online].

Available:   [http://www.ncbi.nlm.nih.gov/pubmed/26684758http://www.pubmedcentral.nih.](http://www.ncbi.nlm.nih.gov/pubmed/26684758%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4683756%20http://www.ijbnpa.org/content/12/1/159)

[gov/articlerender.fcgi?artid=PMC4683756http://www.ijbnpa.org/content/12/1/159](http://www.ncbi.nlm.nih.gov/pubmed/26684758%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4683756%20http://www.ijbnpa.org/content/12/1/159)

[42]  The    Criteria    Committee    of    the    New    York    Heart    Association,     *Classiﬁcation    of*

*Functional   Capacity   and   Objective   Assessment*,   9th   ed.     Boston,   Mass.:    Little,   Brown

and   Co.,   1994.   [Online].   Available:     [http://professional.heart.org/professional/General/](http://professional.heart.org/professional/General/UCM_423811_Classification-of-Functional-Capacity-and-Objective-Assessment.jsp)

[UCM](http://professional.heart.org/professional/General/UCM_423811_Classification-of-Functional-Capacity-and-Objective-Assessment.jsp)

[423811](http://professional.heart.org/professional/General/UCM_423811_Classification-of-Functional-Capacity-and-Objective-Assessment.jsp)

[Classiﬁcation-of-Functional-Capacity-and-Objective-Assessment.jsp](http://professional.heart.org/professional/General/UCM_423811_Classification-of-Functional-Capacity-and-Objective-Assessment.jsp)

[43]  S.  L.  Carroll,  K.  Harkness,  and  M.  H.  Mcgillion,  “A  Comparison  of  the  NYHA  Classiﬁcation

and   the   Duke   Treadmill   Score   in   Patients   with   Cardiovascular   Disease,”   *Open   Journal*

*of  Nursing*,  vol.  4,  pp.  774–783,  2014.  [Online].  Available:   [http://www.scirp.org/journal/](http://www.scirp.org/journal/ojn%20http://dx.doi.org/10.4236/ojn.2014.411083%20http://creativecommons.org/licenses/by/4.0/)

[ojnhttp://dx.doi.org/10.4236/ojn.2014.411083http://creativecommons.org/licenses/by/4.0/](http://www.scirp.org/journal/ojn%20http://dx.doi.org/10.4236/ojn.2014.411083%20http://creativecommons.org/licenses/by/4.0/)

[44]  L.  Goldman,  B.  Hashimoto,  E.  F.  Cook,  and  a.  Loscalzo,  “Comparative  reproducibility  and

validity  of  systems  for  assessing  cardiovascular  functional  class:  advantages  of  a  new  speciﬁc

activity  scale.”  *Circulation*,  vol.  64,  no.  6,  pp.  1227–1234,  1981.

[45]  H.   W.   Christensen,   T.   Haghfelt,   W.   Vach,   A.   Johansen,   and   P.   F.   Hoilund-Carlsen,

“Observer   reproducibility   and   validity   of   systems   for   clinical   classiﬁcation   of   angina

pectoris:     comparison   with   radionuclide   imaging   and   coronary   angiography,”   *Clinical*

*Physiology  and  Functional  Imaging*,  vol.  26,  no.  1,  pp.  26–31,  jan  2006.  [Online].  Available:

<http://doi.wiley.com/10.1111/j.1475-097X.2005.00643.x>

[46]  C.    Raphael,    C.    Briscoe,    J.    Davies,    Z.    Ian    Whinnett,    C.    Manisty,    R.    Sutton,

J.   Mayet,    D.   P.   Francis,    and   C.   Raphael,    “Limitations   of   the   New   York   Heart

Association    functional    classiﬁcation    system    and    self-reported    walking    distances    in

chronic    heart    failure,”     *Heart*,     vol.    93,     pp.    476–482,     2007.    [Online].    Available:

<http://heart.bmj.com/content/heartjnl/93/4/476.full.pdf>

[47]  S.  H.  Kubo,   S.  Schulman,   R.  C.  Starling,   M.  Jessup,   D.  Wentworth,   and  D.  Burkhoﬀ,

“Development   and   validation   of   a   patient   questionnaire   to   determine   New   York   heart

association  classiﬁcation,”  *Journal  of  Cardiac  Failure*,  vol.  10,  no.  3,  pp.  228–235,  2004.

[Online].  Available:    [http://www.sciencedirect.com/science/article/pii/S1071916403007450?](http://www.sciencedirect.com/science/article/pii/S1071916403007450?via%3Dihub)

[via%3Dihub](http://www.sciencedirect.com/science/article/pii/S1071916403007450?via%3Dihub)

[48]  Z.    S.    Nasreddine,    N.    A.    Phillips,    V.    Bedirian,    S.    Charbonneau,    V.    Whitehead,

I.   Collin,   J.   L.   Cummings,   and   H.   Chertkow,   “The   Montreal   Cognitive   Assessment,

MoCA:   A   Brief   Screening   Tool   For   Mild   Cognitive   Impairment,”    *Journal   of   the*

*American  Geriatrics  Society*,   vol.  53,   no.  4,   pp.  695–699,   apr  2005.  [Online].  Available:

<http://doi.wiley.com/10.1111/j.1532-5415.2005.53221.x>

[49]  P.  White,  “An  immersive  virtual  reality  navigational  tool  for  diagnosing  and  treating  neu-

rodegeneration,”  2016.

46

[50]  P.   White,   Z.   Moussavi,   and   P.   White,   “Neurocognitive   Treatment   for   a   Patient   with

Alzheimer’s   Disease   Using   a   Virtual   Reality   Navigational   Environment,”    *Journal   of*

*Experimental  Neuroscience*,  p.  129,  nov  2016.  [Online].  Available:  [http://www.la-press.com/](http://www.la-press.com/neurocognitive-treatment-for-a-patient-with-alzheimers-disease-using-a-article-a6014)

[neurocognitive-treatment-for-a-patient-with-alzheimers-disease-using-a-article-a6014](http://www.la-press.com/neurocognitive-treatment-for-a-patient-with-alzheimers-disease-using-a-article-a6014)

[51]  Fitbit  Inc.,  “Help  article:  What  should  I  know  about  my  heart  rate  data?”   2017.  [Online].

Available:  [https://help.ﬁtbit.com/articles/en](https://help.fitbit.com/articles/en_US/Help_article/1565)

[US/Help](https://help.fitbit.com/articles/en_US/Help_article/1565)

[article/1565](https://help.fitbit.com/articles/en_US/Help_article/1565)

[52]  J.  Allen,  “Photoplethysmography  and  its  application  in  clinical  physiological  measurement,”

*Physiol.  Meas*,  vol.  28,  pp.  1–39,  2007.  [Online].  Available:  [http://iopscience.iop.org/article/](http://iopscience.iop.org/article/10.1088/0967-3334/28/3/R01/pdf)

[10.1088/0967-3334/28/3/R01/pdf](http://iopscience.iop.org/article/10.1088/0967-3334/28/3/R01/pdf)

[53]  R.  Wang,  G.  Blackburn,  M.  Desai,  D.  Phelan,  L.  Gillinov,  P.  Houghtaling,  M.  Gillinov,

C.  MA,  M.  H,  L.  RMT,  T.  DJ,  E.-A.  F,  and  P.  MS,  “Accuracy  of  Wrist-Worn  Heart  Rate

Monitors,”  *JAMA  Cardiology*,  vol.  313,  no.  6,  pp.  625–626,  oct  2016.  [Online].  Available:

<http://cardiology.jamanetwork.com/article.aspx?doi=10.1001/jamacardio.2016.3340>

[54]  Fitbit  Inc.,  “Help  article:  How  does  my  Fitbit  device  count  steps?”  2017.  [Online].  Available:

[https://help.ﬁtbit.com/articles/en](https://help.fitbit.com/articles/en_US/Help_article/1143)

[US/Help](https://help.fitbit.com/articles/en_US/Help_article/1143)

[article/1143](https://help.fitbit.com/articles/en_US/Help_article/1143)

[55]  ——,   “Help   article:     How   accurate   are   Fitbit   devices?”      2017.   [Online].   Available:

[https://help.ﬁtbit.com/articles/en](https://help.fitbit.com/articles/en_US/Help_article/1136)

[US/Help](https://help.fitbit.com/articles/en_US/Help_article/1136)

[article/1136](https://help.fitbit.com/articles/en_US/Help_article/1136)

[56]  K.  M.  Diaz,  D.  J.  Krupka,  M.  J.  Chang,  J.  Peacock,  Y.  Ma,  J.  Goldsmith,  J.  E.  Schwartz,

and  K.  W.  Davidson,  “Fitbit?:  An  accurate  and  reliable  device  for  wireless  physical  activity

tracking,”  2015.

[57]  A.     C.     Acta,      C.     Beleites,      U.     Neugebauer,      T.     Bocklitz,      C.     Kraﬀt,      and

J.     Popp,     “Sample     size     planning     for     classiﬁcation     models,”     *Analytica     Chim-*

*ica    Acta*,     vol.    760,     pp.    25–33,     jan    2013.    [Online].    Available:

[https://ac.els-](https://ac.els-cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?_tid=9248c51a-ab1f-11e7-8f70-00000aacb362&acdnat=1507353954_b1a7e3043cc488c46517c991cdb35d61%20http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

[cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?](https://ac.els-cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?_tid=9248c51a-ab1f-11e7-8f70-00000aacb362&acdnat=1507353954_b1a7e3043cc488c46517c991cdb35d61%20http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

[tid=9248c51a-ab1f-](https://ac.els-cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?_tid=9248c51a-ab1f-11e7-8f70-00000aacb362&acdnat=1507353954_b1a7e3043cc488c46517c991cdb35d61%20http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

[11e7-8f70-00000aacb362&acdnat=1507353954](https://ac.els-cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?_tid=9248c51a-ab1f-11e7-8f70-00000aacb362&acdnat=1507353954_b1a7e3043cc488c46517c991cdb35d61%20http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

[b1a7e3043cc488c46517c991cdb35d61http:](https://ac.els-cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?_tid=9248c51a-ab1f-11e7-8f70-00000aacb362&acdnat=1507353954_b1a7e3043cc488c46517c991cdb35d61%20http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

[//www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub](https://ac.els-cdn.com/S0003267012016479/1-s2.0-S0003267012016479-main.pdf?_tid=9248c51a-ab1f-11e7-8f70-00000aacb362&acdnat=1507353954_b1a7e3043cc488c46517c991cdb35d61%20http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub)

[58]  R. L. Figueroa,  Q. Zeng-Treitler,  S. Kandula,  and L. H. Ngo,  “Predicting sample size required

for  classiﬁcation  performance,”   *BMC  Medical  Informatics  and  Decision  Making*,   vol.  12,

no.  1,   p.  8,   dec  2012.  [Online].  Available:   [http://bmcmedinformdecismak.biomedcentral.](http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-8)

[com/articles/10.1186/1472-6947-12-8](http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-8)

[59]  J.  Brownlee,  “How  Much  Training  Data  is  Required  for  Machine  Learning?”  jul  2017.  [On-

line]. Available:  [https://machinelearningmastery.com/much-training-data-required-machine-](https://machinelearningmastery.com/much-training-data-required-machine-learning/)

[learning/](https://machinelearningmastery.com/much-training-data-required-machine-learning/)

[60]  E.  Seto,  K.  J.  Leonard,  C.  Masino,  J.  A.  Cafazzo,  J.  Barnsley,  and  H.  J.  Ross,  “Attitudes  of

heart  failure  patients  and  health  care  providers  towards  mobile  phone-based  remote  monitor-

ing,”  *Journal  of  Medical  Internet  Research*,  vol.  12,  no.  4,  pp.  3–12,  2010.

[61]  Analytics  Vidhya  Content  Team,  “Tutorial  on  5  Powerful  R  Packages  used  for  imputing

missing  values,”  2016.  [Online].  Available:  [https://www.analyticsvidhya.com/blog/2016/03/](https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/)

[tutorial-powerful-packages-imputing-missing-values/](https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/)

47

[62]  D.    J.    Stekhoven,    “Package    ’missForest’,”    Tech.    Rep.,    2016.    [Online].    Available:

<https://cran.r-project.org/web/packages/missForest/missForest.pdf>

[63]  H.   Tanaka,   K.   D.   Monahan,   and   D.   R.   Seals,   “Age-Predicted   Maximal   Heart   Rate

Revisited,”  2001.  [Online].  Available:  [http://www.onlinejacc.org/content/accj/37/1/153.full.](http://www.onlinejacc.org/content/accj/37/1/153.full.pdf?download=false)

[pdf?download=false](http://www.onlinejacc.org/content/accj/37/1/153.full.pdf?download=false)

[64]  Medscape,  “Sectral,  (acebutolol)  dosing,  indications,  interactions,  adverse  eﬀects,  and  more.”

[Online].  Available:  <https://reference.medscape.com/drug/sectral-acebutolol-342354>

[65]  D.  F.  McAuley,   “Beta  Blockers  -  common  dosage  guidelines,”   2017.  [Online].  Available:

<http://www.globalrph.com/beta.htm>

[66]  Heart   and   Stroke   Foundation,   “Beta-blockers   —   Heart   and   Stroke   Foundation,”   2017.

[Online].   Available:      [https://www.heartandstroke.ca/heart/treatments/medications/beta-](https://www.heartandstroke.ca/heart/treatments/medications/beta-blockers)

[blockers](https://www.heartandstroke.ca/heart/treatments/medications/beta-blockers)

[67]  D.  Jurafsky  and  J.  Martin,  “Hidden  Markov  Models,”  in  *Speech  and  Language  Processing*,

3rd   ed.     Pearson,   2017,   ch.   9,   p.   21.   [Online].   Available:    [https://web.stanford.edu/$*\*](https://web.stanford.edu/$\sim%20$jurafsky/slp3/9.pdf%20https://web.stanford.edu/$\sim%20$jurafsky/slp3/)*−*

[sim$jurafsky/slp3/9.pdfhttps://web.stanford.edu/$*\*sim$jurafsky/slp3/](https://web.stanford.edu/$\sim%20$jurafsky/slp3/9.pdf%20https://web.stanford.edu/$\sim%20$jurafsky/slp3/)

[68]  A.   Bobick,   I.   Essa,   A.   Chakraborty,   and   Udacity,   “Markov   Models,”   2015.   [Online].

Available:  <https://www.youtube.com/watch?v=4XqWadvEj2k>

[69]  P.  A.  Gagniuc,  *Markov  chains:  from  theory  to  implementation  and  experimentation*,  1st  ed.

John  Wiley  and  Sons,  Inc,  2017.

[70]  J.  O’Connell  and  S.  Højsgaard,   “Hidden  Semi  Markov  Models  for  Multiple  Observation

Sequences:   The  mhsmm  Package  for  R,”  *Journal  of  Statistical  Software*,   vol.  39,   no.  4,

pp.  1–22,   2011.  [Online].  Available:    [http://www.jstatsoft.org/index.php/jss/article/view/](http://www.jstatsoft.org/index.php/jss/article/view/v039i04/v39i04.pdf%20http://www.jstatsoft.org/v39/i04/)

[v039i04/v39i04.pdfhttp://www.jstatsoft.org/v39/i04/](http://www.jstatsoft.org/index.php/jss/article/view/v039i04/v39i04.pdf%20http://www.jstatsoft.org/v39/i04/)

[71]  L.   Rabiner,   “A   tutorial   on   hidden   Markov   models   and   selected   applications   in   speech

recognition,”  *Proceedings  of  the  IEEE*,  vol.  77,  no.  2,  pp.  257–286,  1989.  [Online].  Available:

<http://ieeexplore.ieee.org/document/18626/>

[72]  A.  Bobick,  I.  Essa,  A.  Chakraborty,  and  Udacity,  “Hidden  Markov  Models,”  2015.  [Online].

Available:  <https://www.youtube.com/watch?v=5araDjcBHMQ>

[73]  R.   M.   Altman   and   R.   Mackay   Altman,   “Mixed   Hidden   Markov   Models   Mixed   Hidden

Markov   Models:     An   Extension   of   the   Hidden   Markov   Model   to   the   Longitudinal

Data   Setting,”   *Journal   of   the   American   Statistical   Association*,   vol.   102477,   pp.   201–

210,   2007.  [Online].  Available:    [http://amstat.tandfonline.com/action/journalInformation?](http://amstat.tandfonline.com/action/journalInformation?journalCode=uasa20%20http://dx.doi.org/10.1198/016214506000001086)

[journalCode=uasa20http://dx.doi.org/10.1198/016214506000001086](http://amstat.tandfonline.com/action/journalInformation?journalCode=uasa20%20http://dx.doi.org/10.1198/016214506000001086)

[74]  T.   Mailund   and   C.   N.   Storm   Pedersen,   “Machine   Learning   in   Bioinformatics   Lecture

Week   5   -   Hidden   Markov   Models   Selecting   model   parameters   or   “training”   Hidden

Markov   Models,”   Aarhus,   Denmark,   p.   56,   2014.   [Online].   Available:     [http://users-](http://users-birc.au.dk/cstorm/courses/MLiB_f14/slides/hidden-markov-models-4.pdf)

[birc.au.dk/cstorm/courses/MLiB](http://users-birc.au.dk/cstorm/courses/MLiB_f14/slides/hidden-markov-models-4.pdf)

[f14/slides/hidden-markov-models-4.pdf](http://users-birc.au.dk/cstorm/courses/MLiB_f14/slides/hidden-markov-models-4.pdf)

48

[75]  B.   Jelinek,   “Review   on   Training   Hidden   Markov   Models   with   Multiple   Observations.”

[Online].   Available:     [https://www.isip.piconepress.com/courses/msstate/ece](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[8443/papers/](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[2001](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[spring/multi](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[obs/p00](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[paper](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[v0.pdf](https://www.isip.piconepress.com/courses/msstate/ece_8443/papers/2001_spring/multi_obs/p00_paper_v0.pdf)

[76]  User34790,  R.  de  Azevdeo,  Morat,  Hxd1011,  Y.  Bulatov,  Masterfool,  and  F.  Dernoncourt,

“What  is  the  diﬀerence  between  the  forward-backward  and  Viterbi  algorithms?    -  Cross

Validated,”   2016.   [Online].   Available:    [https://stats.stackexchange.com/questions/31746/](https://stats.stackexchange.com/questions/31746/what-is-the-difference-between-the-forward-backward-and-viterbi-algorithms)

[what-is-the-diﬀerence-between-the-forward-backward-and-viterbi-algorithms](https://stats.stackexchange.com/questions/31746/what-is-the-difference-between-the-forward-backward-and-viterbi-algorithms)

[77]  L.  J.  Rodr´ıguez  and  I.  Torres,  “Comparative  Study  of  the  Baum-Welch  and  Viterbi  Training

Algorithms  Applied  to  Read  and  Spontaneous  Speech  Recognition,”  in  *Pattern  Recognition*

*and  Image  Analysis*.    Springer,  Berlin,  Heidelberg,  2003,  pp.  847–857.  [Online].  Available:

[http://link.springer.com/10.1007/978-3-540-44871-6](http://link.springer.com/10.1007/978-3-540-44871-6_98)

[98](http://link.springer.com/10.1007/978-3-540-44871-6_98)

[78]  J.  O’Connell  and  S.  Højsgaard,  “Package   mhsmm  ’,”  *CRAN*,  no.  0.4.16,  2017.

[79]  S.   Sayad,   “Model   Evaluation   -   Classiﬁcation,”   p.   1.   [Online].   Available:    [http://chem-](http://chem-eng.utoronto.ca/$\sim%20$datamining/dmc/model_evaluation_c.htm)

[eng.utoronto.ca/$*\*sim$datamining/dmc/model](http://chem-eng.utoronto.ca/$\sim%20$datamining/dmc/model_evaluation_c.htm)

[evaluation](http://chem-eng.utoronto.ca/$\sim%20$datamining/dmc/model_evaluation_c.htm)

[c.htm](http://chem-eng.utoronto.ca/$\sim%20$datamining/dmc/model_evaluation_c.htm)

[80]  MedCalc,    “ROC   curve   analysis   with   MedCalc,”    2017.   [Online].   Available:      [https:](https://www.medcalc.org/manual/roc-curves.php)

[//www.medcalc.org/manual/roc-curves.php](https://www.medcalc.org/manual/roc-curves.php)

[81]  J.  Davis  and  M.  Goadrich,  “The  Relationship  Between  Precision-Recall  and  ROC  Curves,”

*ICML  ’06  Proceedings  of  the  23rd  international  conference  on  Machine  learning*,  pp.  233–240,

2006.  [Online].  Available:  <https://dl.acm.org/citation.cfm?id=1143874>

[82]  P.   A.   Flach,   “ICML’04   tutorial   on   ROC   analysis,”   p.   3,   2004.   [Online].   Available:

[http://www.cs.bris.ac.uk/$*\*sim$ﬂach/ICML04tutorial/](http://www.cs.bris.ac.uk/$\sim%20$flach/ICML04tutorial/)

[83]  Gung  and  Dsimcha,  “ROC  vs  precision-and-recall  curves,”  2013.  [Online].  Available:  [https:](https://stats.stackexchange.com/questions/7207/roc-vs-precision-and-recall-curves/7210#7210)

[//stats.stackexchange.com/questions/7207/roc-vs-precision-and-recall-curves/7210#7210](https://stats.stackexchange.com/questions/7207/roc-vs-precision-and-recall-curves/7210#7210)

[84]  M.   Sokolova   and   G.   Lapalme,   “A   systematic   analysis   of   performance   measures   for

classiﬁcation    tasks,”    *Information    Processing    and    Management*,    vol.    45,    pp.    427–

437,   2009.   [Online].   Available:    [http://rali.iro.umontreal.ca/rali/sites/default/ﬁles/publis/](http://rali.iro.umontreal.ca/rali/sites/default/files/publis/SokolovaLapalme-JIPM09.pdf)

[SokolovaLapalme-JIPM09.pdf](http://rali.iro.umontreal.ca/rali/sites/default/files/publis/SokolovaLapalme-JIPM09.pdf)

[85]  G.   Forman   and   M.   Scholz,   “Apples-to-Apples   in   Cross-Validation   Studies:    Pitfalls   in

Classiﬁer  Performance  Measurement,”  *SIGKDD  Explorations*,  vol.  12,  no.  1,  pp.  49–57,  2010.

[Online].  Available:  [http://www.kdd.org/exploration](http://www.kdd.org/exploration_files/v12-1-p49-forman-sigkdd.pdf)

[ﬁles/v12-1-p49-forman-sigkdd.pdf](http://www.kdd.org/exploration_files/v12-1-p49-forman-sigkdd.pdf)

[86]  V.  Van  Asch,  “Macro-  and  micro-averaged  evaluation  measures,”  2013.  [Online].  Available:

[http://www.cnts.ua.ac.be/$*\*sim$vincent/pdf/microaverage.pdf](http://www.cnts.ua.ac.be/$\sim%20$vincent/pdf/microaverage.pdf)

[87]  J. E. Fieldsend and R. M. Everson, “Visualisation of multi-class ROC surfaces,” in *Proceedings*

*of the ICML 2005 workshop on ROC Analysis in Machine Learning*, 2005, pp. 49 – 56. [Online].

Available:  [http://users.dsic.upv.es/$*\*sim$ﬂip/ROCML2005/papers/ﬁeldsend2CRC.pdf](http://users.dsic.upv.es/$\sim%20$flip/ROCML2005/papers/fieldsend2CRC.pdf)

[88]  F.  J.  Anscombe,  “Graphs  in  Statistical  Analysis,”  *The  American  Statistician*,  vol.  27,  no.  1,

pp.  17–21,  1973.  [Online].  Available:  <http://www.jstor.org/stable/2682899>

49

[89]  X.   Robin,   N.   Turck,   A.   Hainard,   N.   Tiberti,   F.   Lisacek,   J.-C.   Sanchez,   M.   M¨uller,

and    S.    Siegert,    “Package    ’pROC’,”    *CRAN*,    no.    1.10,

2017.    [Online].    Available:

<https://cran.r-project.org/web/packages/pROC/pROC.pdf>

[90]  Y.  Yan,  “Package  ’MLmetrics’,”  *CRAN*,  no.  1.1.1,  2016.  [Online].  Available:  [https://cran.r-](https://cran.r-project.org/web/packages/MLmetrics/MLmetrics.pdf)

[project.org/web/packages/MLmetrics/MLmetrics.pdf](https://cran.r-project.org/web/packages/MLmetrics/MLmetrics.pdf)

[91]  D.   J.   Hand   and   R.   J.   Till,   “A   Simple   Generalisation   of   the   Area   Under   the   ROC

Curve  for  Multiple  Class  Classiﬁcation  Problems,”  *Machine  Learning*,  vol.  45,  no.  2,  pp.

171–186,  2001.  [Online].  Available:  [http://link.springer.com/10.1023/A:1010920819831https:](http://link.springer.com/10.1023/A:1010920819831%20https://link.springer.com/content/pdf/10.1023%2FA%3A1010920819831.pdf)

[//link.springer.com/content/pdf/10.1023%2FA%3A1010920819831.pdf](http://link.springer.com/10.1023/A:1010920819831%20https://link.springer.com/content/pdf/10.1023%2FA%3A1010920819831.pdf)

[92]  R.  Hyndman,  Chl,  A.  W,  Garak,  and  A.  Grigorev,  “How  to  plot  ROC  curves  in  multiclass

classiﬁcation?”  jul  2010.  [Online].  Available:  <https://stats.stackexchange.com/a/32101>

[93]  T.  Kautz,  B.  M.  Eskoﬁer,  and  C.  F.  Pasluosta,  “Generic  performance  measure  for  multiclass-

classiﬁers,”   *Pattern   Recognition*,   vol.   68,   pp.   111–125,   2017.   [Online].   Available:    [https:](https://ac.els-cdn.com/S0031320317301073/1-s2.0-S0031320317301073-main.pdf?_tid=be702404-bf55-11e7-a7f8-00000aab0f6b&acdnat=1509576245_ed9d4621d0296ad7dd85a6839296e0e1)

[//ac.els-cdn.com/S0031320317301073/1-s2.0-S0031320317301073-main.pdf?](https://ac.els-cdn.com/S0031320317301073/1-s2.0-S0031320317301073-main.pdf?_tid=be702404-bf55-11e7-a7f8-00000aab0f6b&acdnat=1509576245_ed9d4621d0296ad7dd85a6839296e0e1)

[tid=be702404-](https://ac.els-cdn.com/S0031320317301073/1-s2.0-S0031320317301073-main.pdf?_tid=be702404-bf55-11e7-a7f8-00000aab0f6b&acdnat=1509576245_ed9d4621d0296ad7dd85a6839296e0e1)

[bf55-11e7-a7f8-00000aab0f6b&acdnat=1509576245](https://ac.els-cdn.com/S0031320317301073/1-s2.0-S0031320317301073-main.pdf?_tid=be702404-bf55-11e7-a7f8-00000aab0f6b&acdnat=1509576245_ed9d4621d0296ad7dd85a6839296e0e1)

[ed9d4621d0296ad7dd85a6839296e0e1](https://ac.els-cdn.com/S0031320317301073/1-s2.0-S0031320317301073-main.pdf?_tid=be702404-bf55-11e7-a7f8-00000aab0f6b&acdnat=1509576245_ed9d4621d0296ad7dd85a6839296e0e1)

[94]  G.  Tsoumakas,  I.  Katakis,  and  I.  Vlahavas,  “Mining  Multi-label  Data,”  in  *Data  Mining  and*

*Knowledge  Discovery  Handbook*.    Boston,  MA:  Springer  US,  2009,  pp.  667–685.  [Online].

Available:  [http://link.springer.com/10.1007/978-0-387-09823-4](http://link.springer.com/10.1007/978-0-387-09823-4_34)

[34](http://link.springer.com/10.1007/978-0-387-09823-4_34)

[95]  F.  Sebastiani,   “Machine  Learning  in  Automated  Text  Categorization,”   *ACM  Computing*

*Surveys*,  vol.  34,  no.  1,  pp.  1–47,  2002.  [Online].  Available:  [http://nmis.isti.cnr.it/sebastiani/](http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf)

[Publications/ACMCS02.pdf](http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf)

[96]  A.  Rehman,  K.  Javed,  and  H.  Babri,  “Feature  selection  based  on  a  normalized  diﬀerence

measure  for  text  classiﬁcation,”  *Information  Processing  and  Management*,  vol.  53,  no.  2,

pp.  473–489,   2017.  [Online].  Available:   [http://www.sciencedirect.com/science/article/pii/](http://www.sciencedirect.com/science/article/pii/S0306457316307051)

[S0306457316307051](http://www.sciencedirect.com/science/article/pii/S0306457316307051)

[97]  J.  Du,   J.  Xu,   H.  Song,   X.  Liu,   and  C.  Tao,   “Optimization  on  machine  learning  based

approaches  for  sentiment  analysis  on  HPV  vaccines  related  tweets,”  *Journal  of  Biomedical*

*Semantics*,  vol.  8,  no.  9,  2017.  [Online].  Available:  [https://jbiomedsem.biomedcentral.com/](https://jbiomedsem.biomedcentral.com/track/pdf/10.1186/s13326-017-0120-6?site=jbiomedsem.biomedcentral.com)

[track/pdf/10.1186/s13326-017-0120-6?site=jbiomedsem.biomedcentral.com](https://jbiomedsem.biomedcentral.com/track/pdf/10.1186/s13326-017-0120-6?site=jbiomedsem.biomedcentral.com)

[98]  F.  Kayaalp,  A.  Zengin,  R.  Kara,  and  S.  Zavrak,  “Leakage  detection  and  localization  on

water  transportation  pipelines:   a  multi-label  classiﬁcation  approach,”   *Neural  Computing*

*and    Applications*,    vol.    28,    no.    10,    pp.    2905–2914,    oct    2017.    [Online].    Available:

<http://link.springer.com/10.1007/s00521-017-2872-4>

[99]  S.  Wang,  J.  Wang,  Z.  Wang,  and  Q.  Ji,  “Enhancing  multi-label  classiﬁcation  by  modeling

dependencies  among  labels,”  *Pattern  Recognition*,  vol.  47,  no.  10,  pp.  3405–3413,  oct  2014.

[Online].  Available:  <http://linkinghub.elsevier.com/retrieve/pii/S003132031400154X>

[100]  C.    D.    Manning,    P.    Raghavan,    and    H.    Sch¨utze,    “Text    Classiﬁcation    and    Naive

Bayes,”   in   *Introduction   to   Information   Retrieval*.     Cambridge:     Cambridge   University

Press,    2009,    ch.   13,    pp.   253–288.   [Online].   Available:     [https://nlp.stanford.edu/IR-](https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf)

[book/pdf/irbookonlinereading.pdf](https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf)

50

[101]  T.   Segaran,   *Programming   collective   intelligence   :    building   smart   web   2.0   applications*.

O’Reilly,  2007.

[102]  M.  J.  Sandel,  *Justice:  What’s  the  Right  Thing  to  Do?   Harvard  Justice*.    Farrar,  Straus  and

Giroux,  2010.  [Online].  Available:   [http://justiceharvard.org/justice-whats-the-right-thing-](http://justiceharvard.org/justice-whats-the-right-thing-to-do/)

[to-do/](http://justiceharvard.org/justice-whats-the-right-thing-to-do/)

51