**Principles of medical claims prediction:**

1. Let’s first tackle reliably capturing the 1% or the 5% that make up 20-50% of all medical claims cost.
   1. When we get into the mindset of capturing these members, model design features like lookup tables start to make a lot of sense. (See below principle regarding lookup tables).
   2. Also, we begin to use fit metrics that may be more appropriate. (See below regarding conservative predictions).
2. Machine learning models are the hatchet, the lookup tables are the scalpel. Most the value comes from the lookup tables.
   1. Lookup tables are *absolutely* necessary because:
      1. We’re most interested in the sparse observations with “outlier” values.
      2. Sparsity and outliers are not preferable for machine learning models so we do things like capping response variables and grouping sparse predictor categories. There’s no way around it…
      3. Some conditions and drugs that did not exist at the time of model fitting *will* exist at the time of model scoring.
      4. Lookup tables allow us to incorporate human intelligence about what sort of events will precede high claims that it would take a lot of data and proper engineering and model design in order for the model to learn passively.
   2. Milliman uses lookup tables for specialty drugs in its Curv model.
3. It’s OK to have a prediction that is conservative in some circumstances. Use a fit metric that does not reward more for predicting $50K on a $500K member than predicting $1M on that same member.
   1. In evaluation and also in the fitting of the “hatchet” part of the model we would cap the response at 60K for each given month.
4. Deep learning models are not appropriate for medical claims predictive modeling.
   1. Let’s use the example of a 2d cat image with b&w pixels encoded 1 or 0.
      1. We can train a CNN on a series of images that the neural net would begin to learn how to pick up important features like ears or diamond-shaped pupils that represent cats.
      2. Also, we might be able to train with certain sections if the image missing to be able to make the model robust to say, partial images.
      3. Finally we simulate “grainy” images by randomly “deleting” 50% sections of some pixels and develop methods to ensure the model can work under such circumstances.
         1. Perhaps we’d have to interpolate before running the data through the model. E.g. if surrounding pixels are “1” then the missing pixel is “0”
   2. Now let’s try to think through this in the context of medical claims data. Where we have two dimensions: One is time, and another is category of say, diagnosis code.
      1. Let’s say we have complete data…. Then we might try to train a model with all 80K ICD-10 diagnosis codes. There are some problems here that we didn’t have with the cat data.
         1. Whereas, in the image data there is some indication of the “closeness” of one pixel to another, there is no such readily available structure to diagnosis codes.
         2. The events we’re most interested in are rare. Coding practices vary from provider to provider.
            1. How can we ensure that the right pattern is picked up frequently enough to make it into the deep learning model?
            2. How much data is required to pick this up?
            3. How many resources would it take to fit the right model and are we confident that fitting such a model can be done more efficiently than say, a targeted approach to generating cost assumptions for the highest cost conditions, like hemophilia and transplants.
         3. When *applying* the model we have missingness because our third-party data is not complete. Unlike with the cat image where we can do imputation based on random missingness:
            1. We do not *know* which “pixels” are 0 because they are missing and which are 0 because no provider truly diagnosed a member as having a condition.
            2. With images, all pixels have the same meaning and value on their own and there is a spatial structure to them, so imputation by using neighboring pixels is a reasonable solutions.
            3. Medical missingness from 3rd party data may be structured. E.g. we capture a member’s primary care doctor, but not the specialist that is ordering the highest cost drugs.

This would be like the cat missingness not being random. E.g. we’re missing all the data related to “diamond” pupils or pointy ears.

With the cat data if we knew this data could be missing we could try to train it to work well under such circumstances. With medical data, how do we do that?

My solution is to have independent time and datatype specific models and then a second-tier model to combine them and “rules” to ensure the model behaves appropriately when specific pieces of information are received.

1. Lab data is unique in that it can provide a “true positive” for a prediction of health.
   1. With medical and Rx data, a “missing” can mean an individual is healthy and does not require services or that we don’t capture the member’s provider network.
      1. The individuals that we should theoretically be most confident about their good health are those which we appear to have their complete Rx and medical claims and the claims do not indicate any poor health.
   2. With lab data we also have “missing” where we’re compelled to make an assumption about the extent to which the member is truly healthy.
      1. However, when the data is not missing, we get more detailed information about how healthy a member truly is. E.g. we don’t just know that a member saw their provider for an annual physical, we know that the member has low cholesterol.

**Things to Remember:**

* Smart Specialty Table Generation
  + Probably should find a new name for it
  + Reimbursement codes.com is what the FDB drug table is based on
    - Has listing of drugs active on market by FDB brand name and NDCid’s
    - Biologic appl cd and spcl ben id are the
  + Another website was used to like drugs to HCPCS
  + Used Milliman Debit Manual for Rx debits ($129 per debit).
    - Jason Sciborski is the contact
    - I think it costed Humana about $50 K
  + Milliman table only updated annually…. Need monthly monitoring to detect high cost drugs sooner and develop cost assumptions.
    - Either via Reimbursement Codes.com
    - Or via raw data monitoring.
  + About 25% of med+rx costs are from specialty drugs.
    - 2/3 are filled via specialty pharmacy
    - 1/3 paid for by medical benefit.
    - Cost is driving trend.
* Model flow/design.
  + Separate, independent models by time period and data type. Light GBM models
    - Datatype = IP (has revenue\_cd), OP, and Rx (+lab)
  + Independent predictions used in second and third tiers of model to generate a single member-level prediction.
  + Would like the group level prediction to simply be a rolled-up member-level prediction.
    - This means pushing group-level variables to the member-level models including:
      * Group level match rate for datatypes being fit in the member level model
      * Hist\_months\_avail, group\_type/size, etc.
    - This allows an intuitive place to use our lookup tables as the “minimum” for the model’s prediction.
    - Could make a linear adjustment/calibration to the final group-level prediction if necessary.
  + Group-level pfm\_calibrated estimated and included in member-level models. Also fit a “pfm\_calibrated” model. Final prediction is the ratio of one to the other.
    - We know age/gender/covtype based on the info fed *into* the de-id engine
  + Dates:
    - In production: Identify “data\_thru” date and start countback from there.
    - When fitting model, used a 3-month gap… anything not reported as of three months prior to effective date not included in the model.

**Ideas:**

* Have we tried to get a hold of MIB data?
  + This is data that insurance companies share about life insurance events. I’m guessing it includes:
    - Attending physician statements relating to health.
    - The UW decision
* Are you involved in Exam One (Quest) Health PiQture?
  + Seems like they’re trying to one-up Milliman Intelliscript.
  + Sounds like they have contracts with pharmacy data providers, wonder if they would share that data as part of their relationship with Prognos…. Or at least connect Prognos with those who we would need to contract with.