**Progression Analysis**

**High Performance Athlete Development**

High Performance Sport New Zealand

# Pilot Analysis

1. Cycling
2. Canoe Racing

Original authors: Ben Day

Editors:

Date created: March 2020

Last updated: 25/03/2020

# **Purpose**

HP Athlete Development wish to grow collective understanding of the pathways from development level (youth and junior) towards senior level competition. The role of Intelligence in this piece of work is the ability to utilise performance results to determine the pathways of historical medallists. By investigating the progression of performance in the years leading up to a medal winning performance, HPSNZ and Sport expectations can be aligned closer with facts.

|  |  |
| --- | --- |
| **Why** | Equip HPSNZ and NSOs with facts to understand the varied pathways towards Olympic medal. Deepen understanding of performance pathways from development levels to senior elite levels. |
|  |  |
| **How** | Utilise historical performance results to describe the progression of performance towards elite success. Investigate interactions between genders, disciplines, and ages to refine expectations of developing athletes |
|  |  |
| **What** | Analyse historical competition results (via Gracenote) to test existing theories and intuitions.  Develop methods that are reproducible, adaptable, and future-focussed to enable continual improvement and an evolving collective understanding of performance pathways. |

**Initial engagement**

An initial meeting took place in February 2020. Ken Lynch and Dave Wright from HP Athlete Development conveyed the questions they seek to answer. Ben Day and Chris Rawlings listened and began to interpret the role Intelligence can play in answering these questions.

## NZ Olympic/Paralympic Team Profile

**HPSNZ Board Papers 2013/2014**

Prior to the meeting Ken shared Board papers that outlined the predicted makeup of the 2020 NZ Tokyo Olympic team. These papers described the profile of the predicted team by sport and categorised into first-time v returning Olympians. Also mentioned were the implications of returning Olympians anticipated to retire, along with comparisons of NZ team profile to other Olympic nations.

It was acknowledged that Intelligence can be involved with an updated Board Paper after the Tokyo 2020 Olympics and Paralympic Games. The role of Intelligence will be in gathering data on team makeup and improving on the provided insights, such as:

* Clustering/grouping nations to reflect maturity/capacity (financial, political) of system to enable more useful comparisons
* Analyse “first-time” sports that appear at the Game for the first time.
* Differentiate returning medallists from returning medals (i.e. medal-winning campaigns)

## Performance Pathways

**Development levels to senior levels**

An enduring question for HP Athlete Development relates to the transitions of development to senior competition. With a deeper understanding of the typical timelines and differences between and within sports it is possible to better frame expectations for athlete progressions. By growing this collective knowledge about the NZ system, it will be possible to:

* test expert intuitions,
* identify nuanced detail about sports and sport types, and
* ultimately adjust development pathways to leverage NZ sport system strengths

This work will bring confidence to HP Athlete Development methods through meaningful insights. It will also have important flow-on implications to sports as they better understand their performance landscape, talent pools, and historical progressions.

# **Cycling**

Cycling was chosen as the first sport to conduct these analyses. Cycling is a tier 1 targeted sport and has potential for improvement in HP Athlete Development.

## Plan

*Gracenote* will serve as the initial and primary data source for this analysis. HPSNZ subscribe to a service that provides self-service access to global sport competition results. Missing in these results are round progressions such as heats, semi-finals, and qualifying rounds; however, competition outcomes (final results) are available for all sports dating back to the first Olympic Games in 1896. Competition results up to 2016 Rio Olympics will be used in the first instance. Future analysis may incorporate years 2017-2020 for more recent insights.

Junior World Championships results are available from 2006 and will be used to describe the pathway to Olympic medal.

As a starting point the 2020 Olympics cycling disciplines will be the primary focus. These are listed below. The omnium discipline is treated on an outcome basis where constituent races will not initially be analysed in detail.

* Omnium
* Team Sprint
* Team Pursuit
* Keirin
* Sprint

*Gracenote* data will be used to create a list of Olympic medallists. Pinnacle performance results for these medallists (of which some belong to a medal-winning team) will be used to indicate performance progression in years leading up to their Olympic Games medal. Pinnacle events are considered World Championships except in Olympic years, where this is the pinnacle event. The influence of the Commonwealth Games is clear, but at this stage the corresponding World Championships result will be used as pinnacle result.

## Gathering Data

Through the HPSNZ *Gracenote* portal, certain *QlikView* pages can be shown in a results query interface. Both the sport **medallists** and **pinnacle event** datasets were exported from here.

The simple steps below were followed to produce an export (snapshot) of the results database for the specified sport.

**Medallist dataset**

1. Select **Sport** (e.g. Cycling, Canoe Sprint)
2. Select **Competition** = “Olympic Games”
3. Select **Rank** = 1 and 2 and 3 (medallists only)
4. Select **Team Members** = “Add Team Members” (to include an observation for each individual medallist in team events)
5. Click *Format Output for Excel*
6. Click *Export* from the ? dropdown menu (top RHS corner)

**Pinnacle event dataset**

1. Select **Sport** (e.g. Cycling, Canoe Sprint)
2. Select **Competition** = “Olympic Games”, “Juniors World Championships”, “World Championships” (or other Pinnacle event names)
3. Select all **Rank** options (e.g. 1st, 2nd, … )
4. Select **Team Members** = “Add Team Members” (to include an observation for each individual medallist in team events)
5. Click *Format Output for Excel*
6. Click *Export* from the ? dropdown menu (top RHS corner)

# **Decisions and Assumptions**

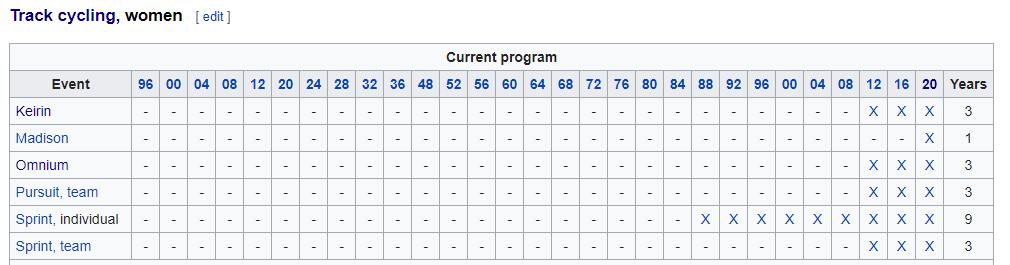
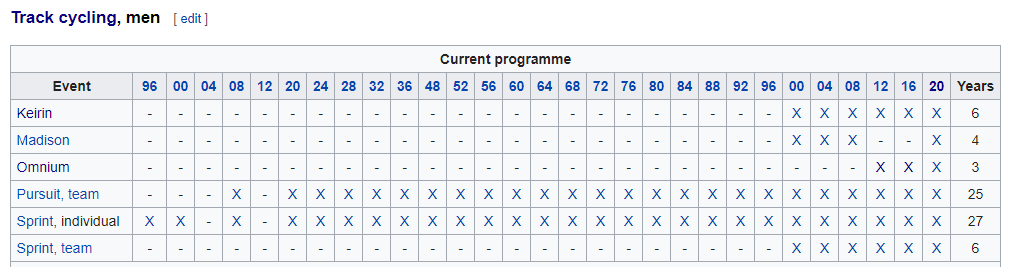
Below are the decisions and assumptions that have been made to focus efforts. Reasoning is provided where appropriate. Some decisions narrow the initial scope but can be returned to in future for wider investigation. In other words, some of these decisions are not final and can be re-visited in future.

1. Initially focus on **first-time medallists only** (exclude repeat medallists – athletes who have medalled at a prior Games)

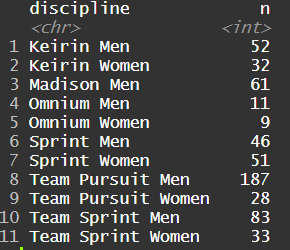
*Why?* *We are most interested in the pathway from junior and development levels to senior success, rather than the repeat success pathway indicative of returning medallists. NOTE first-time medallists may include returning Olympians (i.e. athletes who have been to a prior Games)*

1. Limit medallist dataset to **only disciplines that feature at the Tokyo 2020 Olympic Games**. The first Games to feature Madison, Keirin, and Team Sprint was the Sydney 2000 Olympic Games.

*Why? To narrow our focus to the disciplines that are relevant to track cycling at present. Extending the scope to other disciplines would be useful for crossover analysis between disciplines, but this is not presently a focus of the project.*



1. Initially separate disciplines and **group male and female events together** to find patterns between (types of) disciplines. If disciplines were separated further by gender there would be insufficient data to create individual funnels (see counts below).



*Why? After excluding returning medallists in the same discipline, the number of distinct pinnacle performances amount to a vast majority of male (440) records compared with female (153) records. From the above discipline summary it is observed that 10 or fewer observations would comprise funnel lines for the Omnium Women and Keirin Women disciplines, which is far too few to be useful.*

|  |  |  |
| --- | --- | --- |
| Gender | Number of first time medallist pinnacle performances | Percentage of total (%) |
| Male | 440 | ~74% |
| Female | 153 | ~26% |

*To clarify, these performances do not contain medallists’ performances that are in the same discipline at a subsequent Games. But performances in subsequent Games are included if a medal was won in a different discipline.*

# **Building Phase**

Ken was consulted frequently throughout the tool building phase. This was important to remind us of the purpose of the work and the vision we see in using the insights to assist sports. Some of the refining decision points in this phase are mentioned below.

## Review of Oliver’s Analysis

In 2019 Oliver Stephenson completed an analysis on Junior Athlete Conversions in Track Cycling. *Gracenote* data was manipulated and analysed in R, and can be found in [this repository](https://github.com/HPSNZ/hpad_conversions). This analysis was mathematically thorough and outlined the nature and significance of correlations between Junior World Championship (JWC) attendance and success with Senior level performance.

It was found that more data is needed to form conclusions about the influence of JWC experience on senior medal likelihoods, but the following was found for the existing datasets.

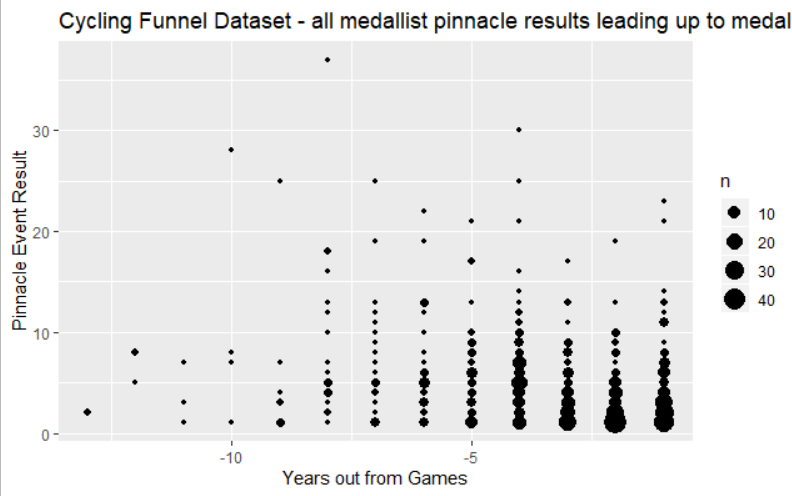
* No evidence to suggest that JWC attendance increases senior medal probability
* 322 of the 468 cyclists (~69%) in the dataset are yet to win their first senior medal. These are effectively “incomplete” data entries… therefore “cycle lengths” only represent some of the underlying dataset.
  + NZ cyclists take more time to win first medal since first JWC appearance
  + NZ cyclists take fewer events to win first medal than other nations (except AUS and GER).
  + “Cycle length” can be misleading at times because of the above

Using these findings and limitations we can narrow the focus of this project. We seek to connect these data to the trajectories of our existing NZ athletes in order to accurately track and contextualise progress from development to senior level performance.

Based on the above finding that JWC experience does not necessarily increase senior medal probability, there is no need to separate datasets by this category.

## Funnel Dataset

World Championship and Olympic Games results will be used to form a set of ‘pinnacle result’ performance data. For individual and team disciplines, annual pinnacle performances will be retrieved for each individual athlete. Showing these in years leading up to the first Olympic medal appears in a scatter plot below.



From the above plot we can observe the spread of the results (y axis) over years leading up to the medal (timeline on x-axis). Beyond 8 years out from Games we have minimal data points to consider (4 or fewer results). It is therefore sensible to limit our funnel timeline to 8 years out from the Games, which neatly lines up with 2 full Olympic cycles.

The remaining dataset can be summarised as follows. It is clear that in years 4-1 out from Games there is a helpful number of results that we can learn from. For years 5-8 out from Games there are fewer and fewer results to draw from, but it is acceptable to use percentiles to convey their dispersion.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Years out from medal** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Number of records | 135 | 122 | 87 | 103 | 50 | 34 | 23 | 19 |

We will use funnel lines to guide expectations of athlete performance tracking towards Olympic medals. These lines will nominally be calculated for each year out as representing the following **percentiles** – showing the dispersion (spread) of results by year.

10th percentile 90% of results fall above this line

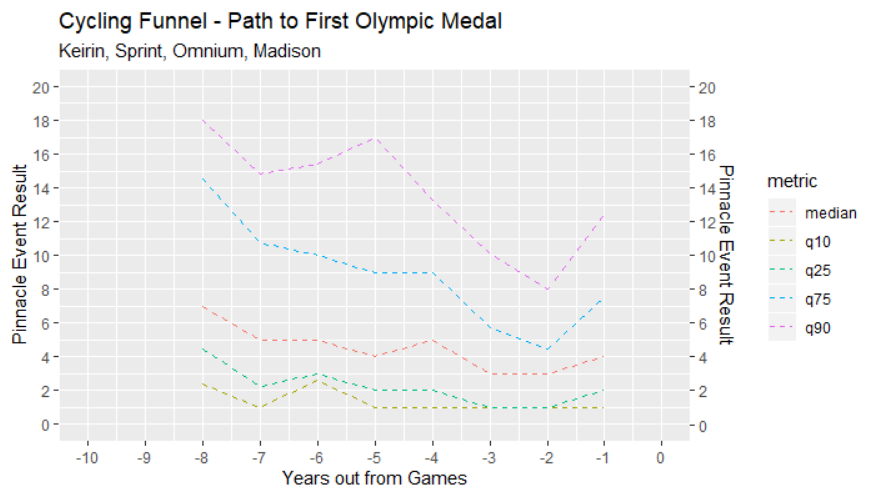
25th percentile (lower quartile) ¾ of results fall above this line

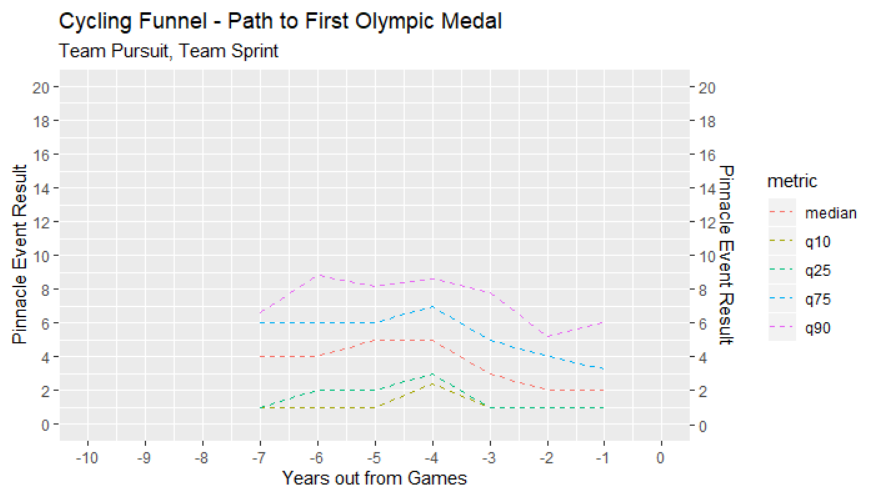
50th percentile (median) represents middle result

75th percentile (upper quartile) ¼ of results fall above this line

90th 10% of results fall above this line

Performance funnels are split by discipline in the first instance. Ideally an output tool will allow the user to select the disciplines to make up the underlying dataset. This is crucial in producing a tool that is specific, customisable and functional for a variety of use cases.

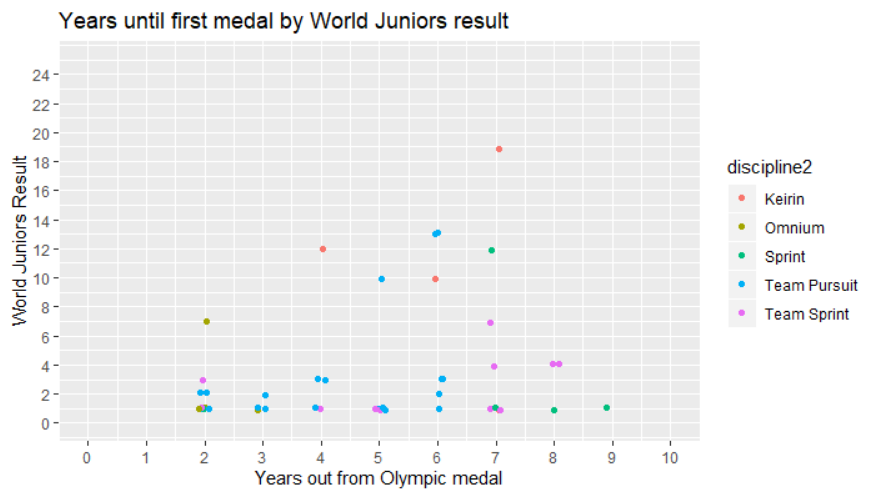




## Emergent questions

1. Constituent funnel lines are made up of prior medallists’ pinnacle event results in the years leading up to their first Olympic medal. How many data points are required for a given year for us to confidently represent the spread of pathways?
2. For which disciplines should we assess progression of race times as well as race rankings? For example, race times in tactical events (such as the Omnium) may not be useful. With a NZ result subset, it appears that 114 of 263 (~43%) of performances have a “Result” field attached to them, which may be used for race time.
3. Of JWC attendees, is there a relationship between result at JWC and time (number of years) until Olympic medal?

Of the subset of Olympic medallists who have attended JWC, there appears to be no relationship between the result achieved at JWC and the length of timeline to an Olympic medal.



# Canoe Racing

As another Tier 1 supported sport within the HPSNZ investment framework it is worth considering the application of this methodology for canoe racing. The ‘timed’ nature of the sport also gives opportunity to analyse the performance progressions of medallists in years leading up to their medal performance.

## Plan

Turning the focus to canoe racing affords the opportunity of testing the original assumptions of this progressions analysis methodology. For the framing of performance progression in canoe racing the following decisions and assumptions are once again made:

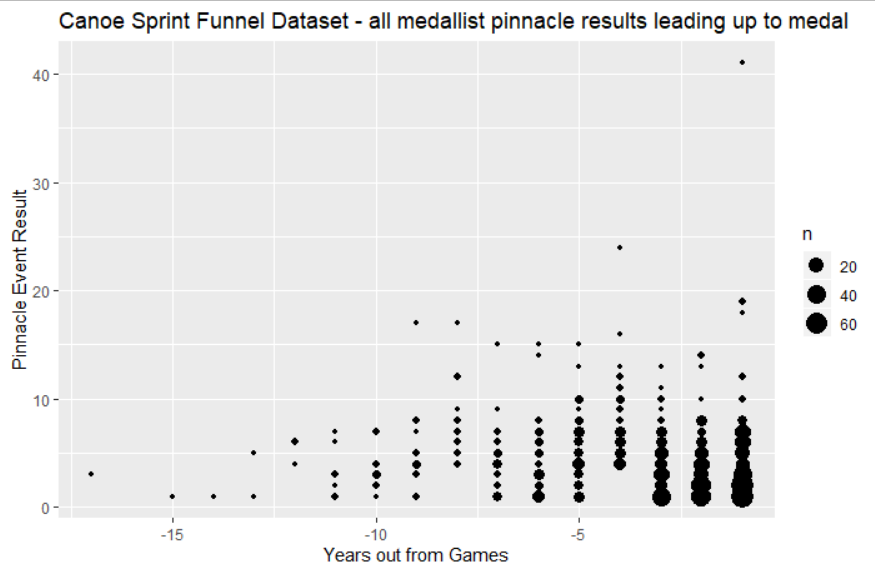
1. First time medallists only (for each discipline)
2. Tokyo 2020 disciplines only
3. Group male and female events together (where necessary)

Ideal context would be provided by a event-specific and gender-specific performance funnel, where NZ athletes would be compared with the most relevant prior medallist data. This is contingent on the depth of the underlying dataset; a judgement must be made on whether to make the separation or not. The following data gathering will describe the process used towards this decision.

## Gathering Data

*Gracenote* was used to extract a) the canoe racing medallist dataset, from which individual performance progression pathways can be derived, and b) an extensive results dataset that contain NZ athlete progressions.

Extracting the medallist dataset (1936-2016) from *Gracenote* provides the following spread of pinnacle performance data. Initial impressions judge that at least 2 cycles of funnel guidelines could be used as potentially as many as 10 years out from Games may be useful.

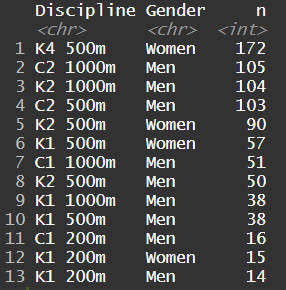


The summary below shows medallist pinnacle performance data by years out from medal-winning performance.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Years out from medal** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **17** |
| Number of records | 274 | 193 | 137 | 63 | 69 | 46 | 28 | 16 | 15 | 12 | 7 | 3 | 2 | 1 | 1 | 1 |

The distribution of pinnacle performance results **by disciplines featured at Tokyo 2020 Games** are shown below and highlight some limitations in the dataset:

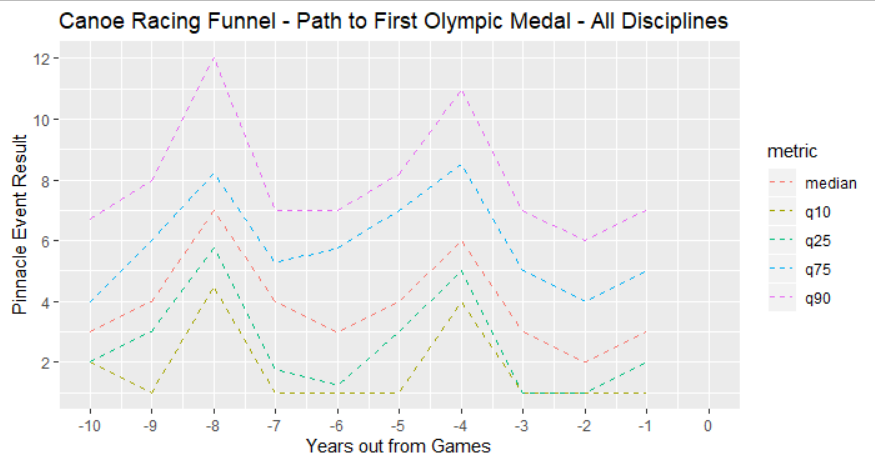
|  |  |
| --- | --- |
| **Discipline** | **Number of pinnacle results** |
| K4 500m | 172 |
| K2 200m | 144 |
| C2 1000m | 105 |
| C2 500m | 105 |
| K2 1000m | 105 |
| K1 500m | 97 |
| C1 1000m | 51 |
| K1 1000m | 39 |
| K1 200m | 33 |
| C1 200m | 17 |



Due to the recent introduction of the 200m distance events there are a limited number of results for these disciplines. When considering the spread of this number (e.g. 17 or 33 results) across a timespan leading up to a Games medal, it is straightforward to see that there may not be enough data at each ‘year out’.

Viewing the distribution of medallist performances by gender reveals the same limitation to 200m events. On this basis it will not be appropriate to produce a separate 200m funnel.

As distinct from the Cycling case study, it appears that the discrepancies between male and female events are not as pronounced with regards numbers of medallist performances. It is desirable then to offer separate male and female event funnels for an interface tool, with the proviso that drawn guidelines would represent ‘enough data’ at each year out from Games. In other words, the funnel tool would be configured to filter out certain years out if there are too few male or female performance results for that year.



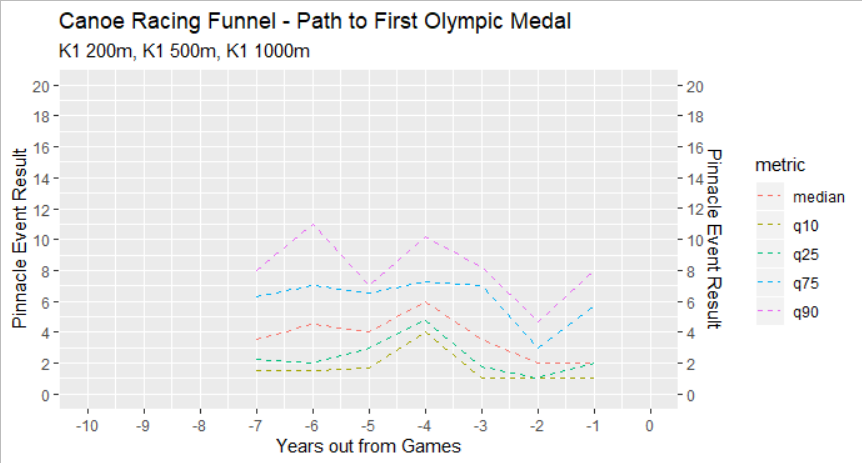
A view of all disciplines together reveals a pattern that highlights the competitiveness of Olympic years. The pinnacle performances of medallists tend to drop by 2-4 places in Olympic years on account of the deeper and tougher competition; medallists’ best results are observed at the Olympiad anti-phase of 2 years out and/or 6 years out from the Games.

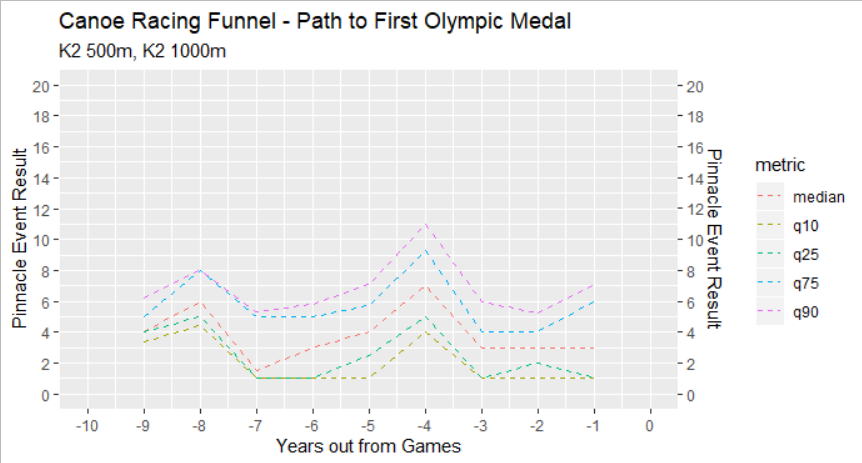
A key observation shows that ¾ of all medallists placed no worse than 5th (i.e. top 5) in the 3 years leading up to the Games medal.

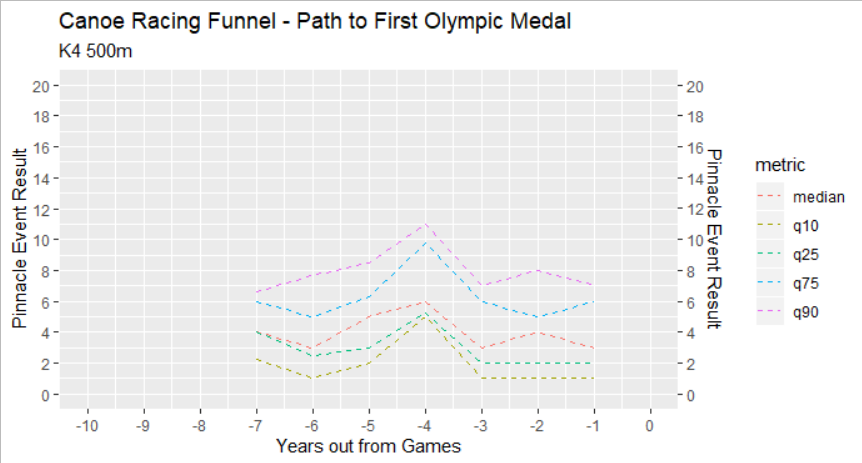
## Grouping by boat type

By grouping the results into disciplines for the same boat type (e.g. C1, C2, K1, K2, K4) we may discover patterns unique or specific to that boat type.

Given that we currently have no canoe athletes in the CRNZ program the canoe disciplines (C1, C2) will be ignored as a grouping at this stage.

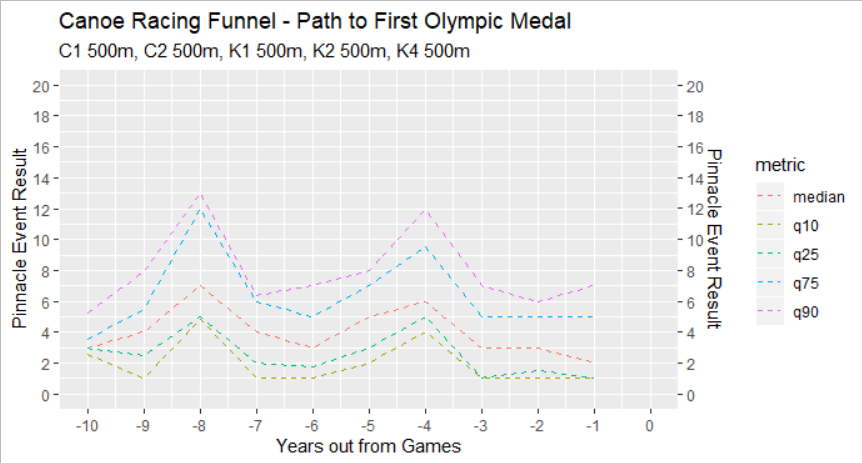


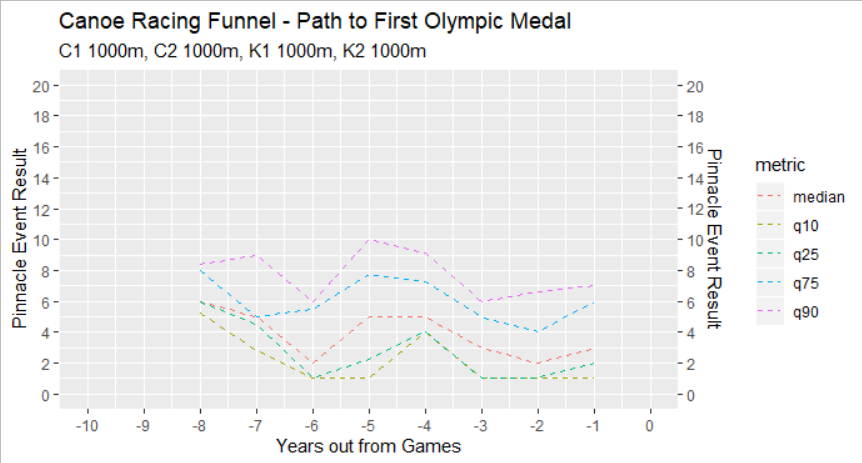




## Grouping by race distance

An alternative grouping method may reveal patterns about races with the same length. Unfortunately, the dataset is too limited for the 200m length race, so we cannot separate the C1 200m and K1 200m events in this way.





## Grouping by gender

