Lecture 11

Monday, November 27, 2023 10:02

BIOS522_Slides11

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Lecture 11:

Survival analysis in clinical trials

Clinical trial phases

- **Pre-clinical**: animal and laboratory testing.
- **Phase 1**: small, first-in-human trials to assess safety.
- **Phase 2**: larger trials to assess safety and early evidence of efficacy.
- **Phase 3**: large trial to assess efficacy and safety to inform licensure.

2

Phase 3 trials

- **Clinically meaningful** primary endpoint, directly measures how a patient:
  - "**Feels**" (symptoms),
  - "**Functions**" (the ability to perform activities in daily life), or
  - "**Survives**"
- *Example*: Idiopathic pulmonary fibrosis
  - Primary endpoints: all-cause mortality, all-cause non-elective hospitalization
  - Indirect (surrogate) endpoints: forced vital capacity, diffusion capacity

3

Intention-to-treat vs. per protocol

- **Intention-to-treat (ITT):** participants are analyzed according to their treatment assignment, not the treatment actually received
  - Time origin = randomization, first dose of study drug
  - Ignores protocol deviations
  - Preferred for clinical trials

- **Per protocol**: includes only participants who completed the study without major protocol violations
  - Time origin might be shifted to last dose (e.g. 3 dose vaccine)

Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial

Maria Rosario Capeding, Ngoc Huu Tran, Sri Rezeki S Hadinegoro, Hussain Imam HJ Muhammad Ismail, Tawee Chotpitayasunondh, Mary Noreen Chua, Chan Quang Luong, Kusnandi Rusmil, Dewa Nyoman Wirawan, Revathy Nallusamy, Punnee Pitisuttithum, Usa Thisyakorn, In-Kyu Yoon, Diane van der Vliet, Edith Langevin, Thelma Laot, Yanee Hutagalung, Carina Frago, Mark Boaz, T Anh Wartel, Nadia G Tornieporth, Melanie Saville, Alain Bouckenooghe, and the CYD14 Study Group*

Goal: **Phase 3**, placebo-controlled trial to assess the efficacy of a 3 dose candidate **dengue vaccine** (injections at 0, 6 and 12 months)

Population: **10,275 children** aged 2-14 years

Outcome variable: The primary outcome was **date of symptom onset** of virologically confirmed clinical dengue

- **Intention-to-treat analysis**: from randomization, received ≥1 injection
- ****Per protocol analysis**: 28 days after 3rd injection

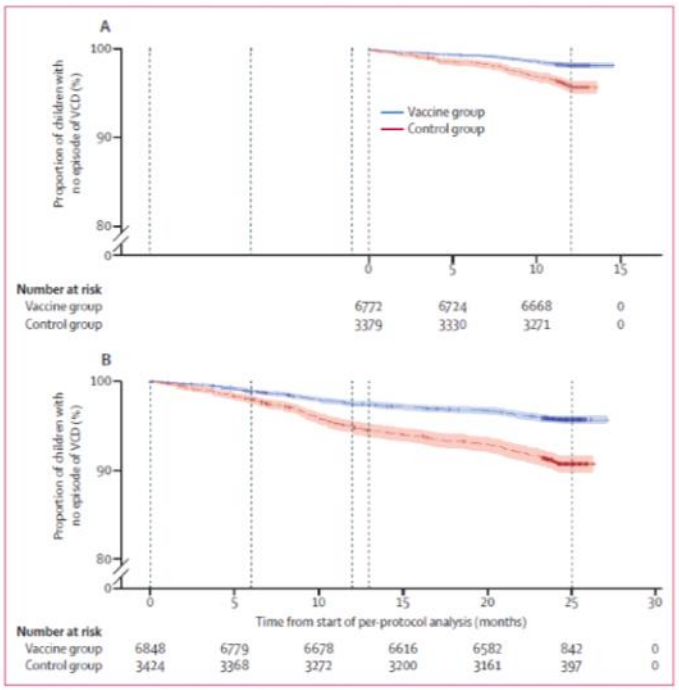


Figure 2: Kaplan-Meier curve for symptomatic virologically-confirmed dengue (VCD) due to any serotypes taking place from 28 days after the third injection (ie, from month 13) in the per-protocol population (A) and at any time during the trial from day 0, irrespective of protocol compliance, in the intention-to-treat population (B)

Per protocol:

- Time origin at 13 months
- 117 cases in 6526 vaccinated PY
- 133 cases in 3227 unvaccinated PY
- VE = 56.5% (95% CI 43.8-66.4%)

Intention-to-treat:

- Time origin at 0 months
- 286 cases in 13571 vaccinated PY
- 309 cases in 6623 unvaccinated PY
- VE = 54.8% (95% CI 46.8-61.7%)

6

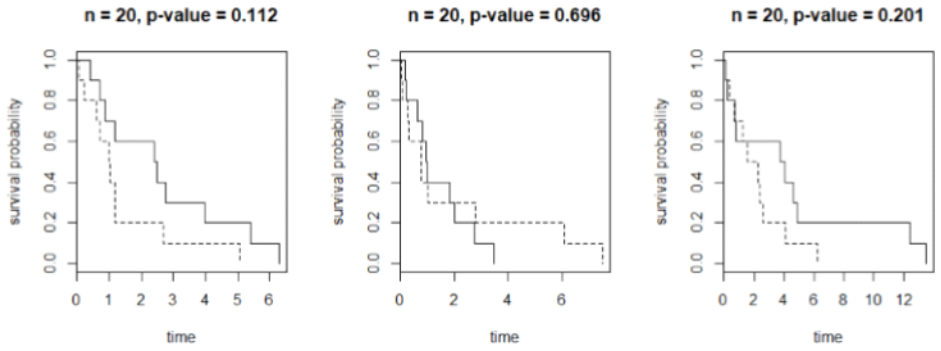
Power and sample size

- Studies must be designed to be able to detect a **clinically meaningfully effect** where one exists
  - Unethical to waste $$, resources, participant time
- **Power** measures our ability to detect a difference between groups (where one exists)
- Power increases with **sample size**
- *A priori*, we must determine the required sample size likely to yield our desired high level of power

7

Example: $\lambda_0 = 0.6$ days^{-1}, HR = 0.5, $n = 20$

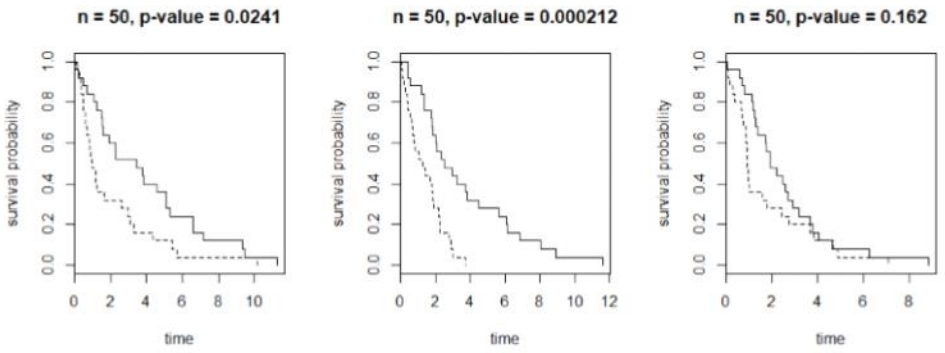
- Even though the true effect is to reduce hazard by 50%, none of the log-rank tests are statistically significant
- Only **34% of the time** will we declare a significant difference



8

Example: $\lambda_0 = 0.6$ days^{-1}, HR = 0.5, $n = 50$

- Even though the true effect is to reduce hazard by 50%, only two of the three log-rank tests are statistically significant
- Only **69% of the time** will we declare a significant difference



9

Example: $\lambda_0 = 0.6$ days^{-1}, HR = 0.5

- We need a sample size of at least $n = 70$ to detect a significant effect 80% of the time
- We need a sample size of at least $n = 90$ to detect a significant effect 90% of the time

Total sample size	Power	Type II error
$n = 20$	34%	66%
$n = 30$	48%	52%
$n = 40$	59%	41%
$n = 50$	69%	31%
$n = 60$	77%	23%
$n = 70$	83%	17%
$n = 80$	87%	13%
$n = 90$	91%	9%
$n = 100$	93%	7%

10

Example: $\lambda_0 = 0.6$ days^{-1}, HR = (0.3, 0.5, 0.7)

- If the effect size is larger (further from H_0: HR = 1), we can achieve the same power with a smaller sample size
- Need $n = 30$ to have 90% power to detect $HR = 0.3$
- Need $n > 100$ to have 90% power to detect $HR = 0.7$

Total sample size	Power $HR = 0.7$	Power $HR = 0.5$	Power $HR = 0.3$
$n = 20$	12%	34%	77%
$n = 30$	16%	48%	91%
$n = 40$	20%	59%	97%
$n = 50$	24%	69%	99%
$n = 60$	28%	77%	>99%
$n = 70$	32%	83%	>99%
$n = 80$	36%	87%	>99%
$n = 90$	39%	91%	>99%
$n = 100$	43%	93%	>99%

11

Sample strategy for power calculations

1. Select the primary endpoint and analysis
2. Determine the desired power (e.g. 80%, 90%)
3. Identify the smallest effect size that you want to be powered to detect
4. Calculate the required number of events
   - *Power directly depends on # of events, not sample size*
5. Establish the incidence rate in the control population
6. Determine the *number of participants to enroll* and the *length of time they should be followed*
7. Adjust for losses to follow-up, etc.

Determining effect size

Example: Suppose we want to detect a 50% improvement in median survival (from 12 months to 18 moths). Assume our data follow an exponential distribution with rate λ_i.

Recall that the median for an exponential is $t_{0.50,\mathrm{i}} = \frac{\log(2)}{\lambda_i}$, so $\lambda_i = \frac{\log(2)}{t_{0.50,\mathrm{i}}}$

Thus, the corresponding hazard ratio is:

$$\theta = \frac{\lambda_1}{\lambda_0} = \frac{\frac{\log(2)}{18 \text{ months}}}{\frac{\log(2)}{12 \text{ months}}} = \frac{12}{18} = 0.667$$

Calculating the number of events

For hazard ratio θ, desired power $1-\beta$, and two-sided level α log rank test, the required total number of events d is approximately:

$$d = \frac{4\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{[\log(\theta)]^2}$$

Example: hazard ratio of 0.7 with 90% power at a 2-sided $\alpha = 0.05$:

$$d = \frac{4(1.96 + 1.282)^2}{[\log(0.7)]^2} \approx 331$$

14

Calculating the total sample size

- Start with a known incidence rate or survival probability in the control group
- If you're starting with a survival probability (e.g. 5 year-survival), assume an exponential distribution in order to calculate the rate λ_0
- Calculate the rate in the intervention group $\lambda_1 = \theta\lambda_0$

- Calculate the expected probability of failure in both the control and intervention groups assuming an exponential distribution
- Calculate the expected number of events $d_1 + d_0$ for a given n

- To increase the number of events, increase n or duration of follow-up
- Adjust for expected attrition rate

15

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Phase 2b Controlled Trial of M72/AS01$_E$ Vaccine to Prevent Tuberculosis

O. Van Der Meeren, M. Hatherill, V. Nduba, R.J. Wilkinson, M. Muyoyeta, E. Van Brakel, H.M. Ayles, G. Henostroza, F. Thienemann, T.J. Scriba, A. Diacon, G.L. Blatner, M.-A. Demoitié, M. Tameris, M. Malahleha, J.C. Innes, E. Hellström, N. Martinson, T. Singh, E.J. Akite, A. Khatoon Azam, A. Bollaerts, A.M. Ginsberg, T.G. Evans, P. Gillard, and D.R. Tait

Goal: Assess safety and efficacy of a **vaccine** for preventing **tuberculosis transmission** from infected individuals. A vaccine targeting latently infected individuals that prevents the development of active (contagious) disease could reduce onward transmission.

16

https://www.nejm.org/doi/full/10.1056/nejmoa1803484

Population: They conducted a **randomized, double-blind, placebo-controlled, Phase 2b trial** (large Phase 2 trial) to evaluate the M72/AS01E vaccine. The trial enrolled **HIV-negative adults** aged 18 to 60 years **with latent M. tuberculosis infection** from 11 sites in Kenya, South Africa, and Zambia.

Participants were randomized in **1:1 ratio** to receive two doses (1 month apart) of experimental vaccine or placebo. Participants were followed for up to 3 years.

Outcome variable: The primary endpoint was **progression to bacteriologically-confirmed active pulmonary TB**, confirmed with sputum sample, not associated with HIV infection. Alternative case definitions were considered as secondary endpoints.

The primary analysis was per protocol, beginning 1 month after the second dose.

17

Predictor variables: Besides trial arm, researchers considered participant sex, country, smoking status, presence of diabetes, age (>25 or ≤25 years), and prior vaccination with the BCG vaccine.

Statistical analysis: Vaccine efficacy was analyzed using a **Cox proportional hazards model** with 90% confidence interval. The log-rank test was used to report a p-value.

The primary end point was met if the lower limit of the two-sided 90% confidence interval for vaccine efficacy against bacteriologically confirmed pulmonary tuberculosis was more than 0%.

A prespecified primary analysis was performed when all the participants had completed **at least 2 years of follow-up**. The final analysis after 3 years of follow-up are not reported in this paper because the data were not yet available.

18

With two-sided 10% significance level, assuming a true vaccine efficacy of 70% (hazard ratio, 30%), they estimated that they needed d=21 cases of pulmonary TB to have 80% power using a log-rank test.

Assuming a mean yearly attack rate of 0.55% in the control group, 2 years of follow-up for each participant, and an attrition rate of 15% over the two-year period, they calculated that 3506 participants were needed.

19

Sample size:

1. Select the primary endpoint and analysis
   - Log-rank test with two-sided significance level (α) of 10%
2. Determine the desired power
   - 80% power
3. Identify the smallest effect size that they want to be powered to detect
   - Vaccine efficacy = 70%/Hazard ratio = 30%
4. Calculate the required number of events
   - $d = 21$ events
5. Establish the incidence rate in the control population
   - Mean yearly attack rate of 0.55% in the control group
6. Determine the *number of participants to enroll* and the *length of time they should be followed*
   - Follow each participant for 2 years
7. Adjust for losses to follow-up, etc.
   - Assuming 15% attrition rate, need $n = 3506$ participants

20

Results: Of 3575 participants who underwent randomization, **3283 were included** in the per protocol analysis. Reasons for exclusion included failure to receive both doses of vaccine or placebo, randomization error, randomization code broken, trial regimen not received according to protocol, non-adherence to the trial regimen schedule, participant did not meet inclusion/exclusion criteria, or development of active TB before the start of the efficacy period.

10 cases of active TB were detected in the vaccine group versus 22 cases in the placebo group. The estimated hazard ratio was 0.46. The **overall vaccine efficacy was 54.0%** (90% CI: 13.9 to 75.4%). The log-rank test p-value was 0.04.

An analysis that used a Cox regression model with adjustment for country, sex, diabetes, age, current smoking status, and previous BCG vaccination gave nearly identical results.

21

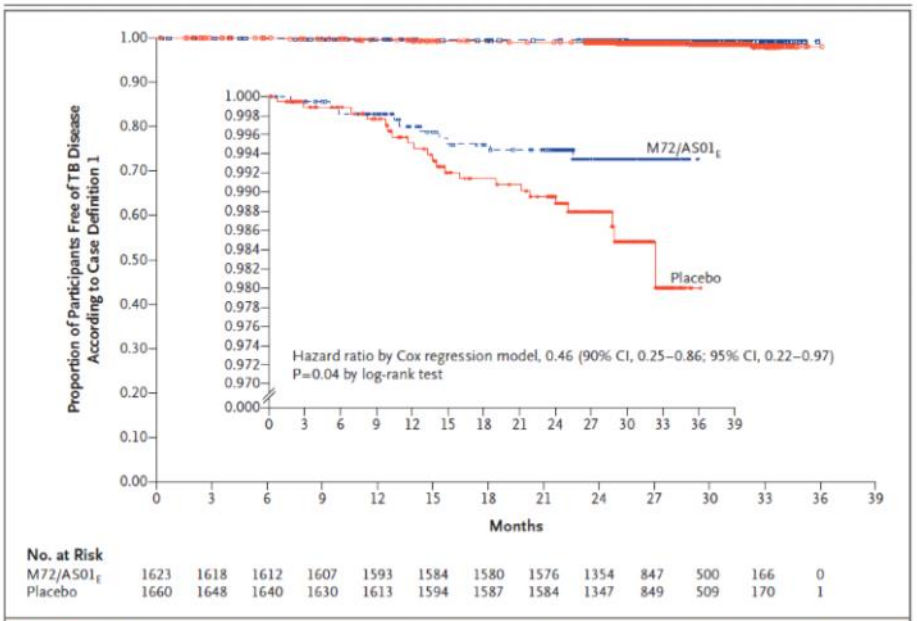


Figure 2. Kaplan–Meier Estimate of Definite Pulmonary Tuberculosis (TB) Disease Not Associated with HIV Infection (First Case Definition).

The analysis was conducted in the according-to-protocol efficacy cohort. The time shown is the time from the beginning of follow-up (i.e., 30 days after dose 2). The inset shows the same data on an enlarged y axis. The decreased number at risk after 24 months reflects the participants for whom follow-up after this time point had not occurred at the date of data lock.

22

Assumptions vs. Reality:

	Assumed	Observed	Impact
Hazard ratio	0.30	0.46	Lower power
Number of events	d=21	d=32	Higher power
Incidence in the control population	0.55%	0.6 per 100 person-years (0.6%)	Similar power
Enrolled	n=3506	n=3568	Similar power
Attrition rate	15% over two years	285/3568 excluded (8%)	Higher power

23

Data monitoring

- Often desirable to conduct **interim analyses** of study data while data collection is ongoing.
  - Ethical: It is wrong to continue to give the inferior treatment to patients.
  - Timely reporting: If the hypothesis of interest has been clearly established, the public may benefit from early reporting.

- Unplanned interim analyses can seriously **inflate type I error**
  - If 1 test is performed at $\alpha = 0.05$, type I error for the trial is 5%.
  - If 2 tests are performed at $\alpha = 0.05$, type I error for the trial is 8.3%.
  - If 3 tests are performed at $\alpha = 0.05$, type I error for the trial is 10.7%.

O'Brien-Fleming approach

- If 2 tests ("looks", interim analyses) are performed during the trial:
  - The first is conducted at $\alpha' = 0.0054$.
  - The second is conducted at $\alpha' = 0.0492$.

- If 3 tests are performed during the trial:
  - The first is conducted at $\alpha' = 0.0006$.
  - The second is conducted at $\alpha' = 0.0151$.
  - The third is conducted at $\alpha' = 0.0471$.

- Both approaches maintain the type I error for the trial at 5%.

Efficacy and safety of sunitinib in patients with advanced gastrointestinal stromal tumour after failure of imatinib: a randomised controlled trial

George D Demetri, Allan T van Oosterom, Christopher R Garrett, Martin E Blackstein, Manisha H Shah, Jaap Verweij, Grant McArthur, Ian R Judson, Michael C Heinrich, Jeffrey A Morgan, Jayesh Desai, Christopher D Fletcher, Suzanne George, Carlo L Bello, Xin Huang, Charles M Baum, Paolo G Casali

Goal: Researchers sought to evaluate the **tolerability** and **anticancer efficacy** of sunitinib for the treatment of unresectable imatinib-resistant gastrointestinal stromal tumors.

Population: They conducted a **multi-center, randomized, double-blinded, placebo-controlled Phase 3** clinical trial. Participants were randomized in a **2:1 ratio** to receive blinded sunitinib or placebo orally once daily in 6-week cycles.

Eligible participants had **histologically proven malignant gastrointestinal stromal tumor** that was not amenable to surgery, radiation, or other approaches, and confirmed failure of previous therapy with imatinib. Participants were enrolled from **56 centers in 11 countries** between December 2003 and January 2005.

https://doi.org/10.1016/S0140-6736(12)61857-1

Outcome variable: The primary endpoint was **time to tumor progression**. Secondary endpoints included progression-free survival and overall survival. The primary efficacy analysis was **intention-to-treat**, including all patients randomized to treatment.

Predictor variables: Besides trial arm, key predictor variables included age, sex, ECOG performance status, histology, tumor burden at baseline (mm), and measurements related to their previous treatment history with imatinib (e.g. duration on therapy).

Statistical analysis: Time to tumor progression was assessed using Kaplan-Meier and compared with the **log-rank test** (primary efficacy analysis). A **stratified log-rank test** and **Cox regression** models were used to explore the potential effects of the stratification factors and patients' baseline characteristics on the primary endpoint.

To determine the **sample size**, researchers described how, at the time the trial was designed (2002-2003), little data was available on expected rates of disease progression in this population who had failed treatment with imatinib. An **informal survey** was done among approximately 25 experts globally. The time to tumor progression after imatinib failure was generally reported to be **less than 4 months**. A **50% improvement** (hazard ratio 0.67) in median time to tumor progression from 4 to 6 months was judged to be clinically meaningful.

28

281 patients with disease progression were estimated to be needed to have **90% power** to detect such an improvement using an unstratified log-rank test with two-sided type I error α=0.05. The investigators estimated that they needed to enroll **357 patients** (238 sunitinib, 119 placebo) to observe 281 patients with progressive disease by the end of the follow-up period.

Pre-specified **interim analyses** were planned after 141 and 211 patients had documented progressive disease. The significance levels at each interim analysis were determined using the **O'Brien-Fleming approach.**

29

<u>Results:</u> The trial was **stopped early** after an interim analysis of the first 149 events showed significantly longer time to tumor progression in patients treated with sunitinib when compared to placebo.

Median time to tumor progression for the ITT population was more than four times as long with sunitinib (27.3 weeks, 95% CI: 16.0-32.1) as with placebo treatment (6.4 weeks, 95% CI: 4.4-10.0) (**hazard ratio 0.33**, 95% CI 0.23-0.47, p<0.0001).

30

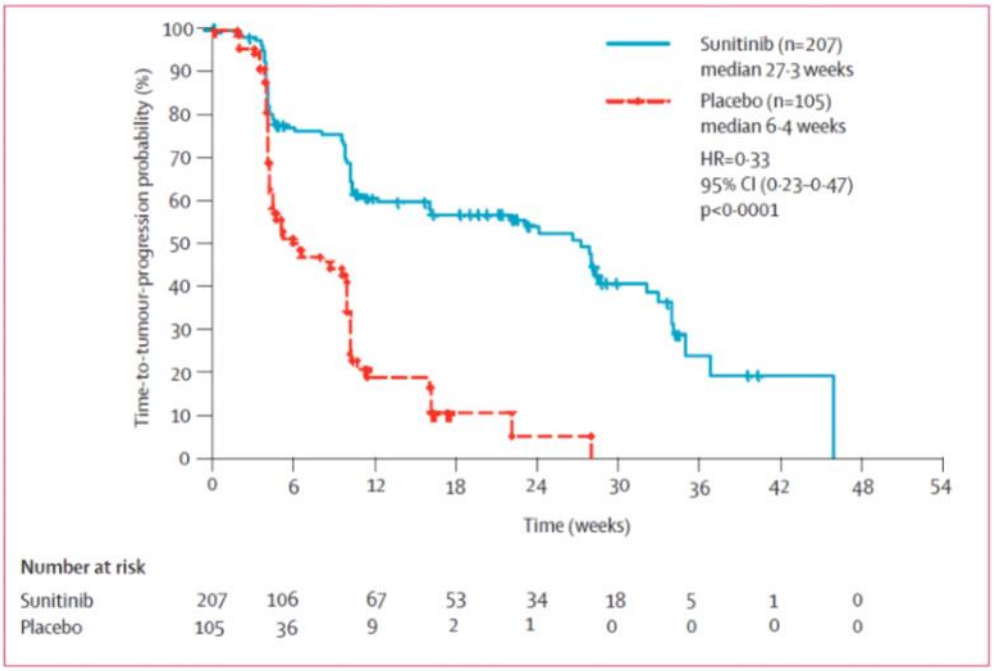


Figure 2: Kaplan-Meier estimates of time to tumour progression
Results represent central radiology assessment of ITT population.

31

Similar results were seen in the stratified analysis and in a Cox proportional hazards regression model. For all subgroups, the hazard ratio was less than 0.5, indicating that all subgroups analyzed benefited from sunitinib therapy compare with placebo.

32

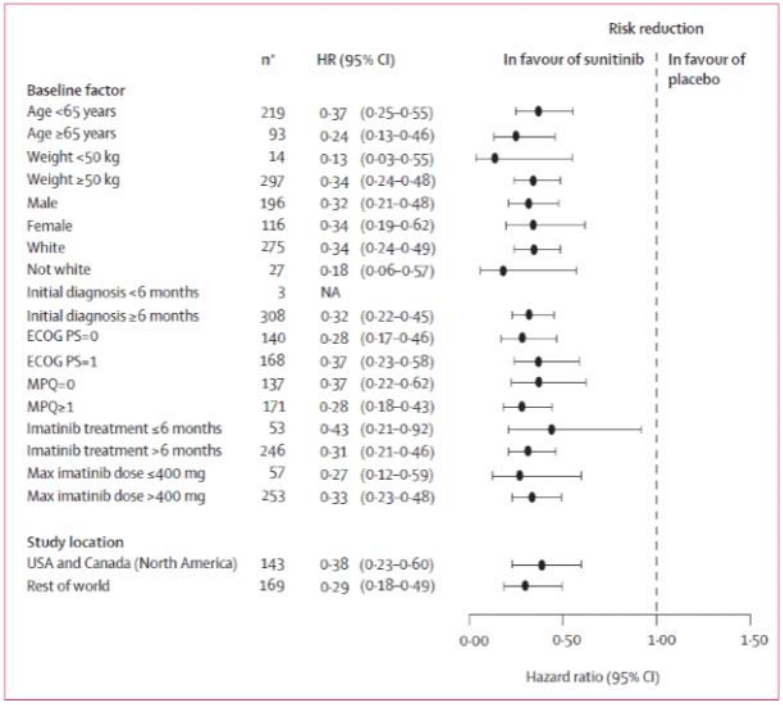


Figure 3: **Cox proportional hazards analysis of time-to-tumour progression treatment comparisons controlling for individual baseline factors**
NA=not applicable. PS=performance status. Max=maximum. MPQ=McGill Pain Questionnaire. *All pairs of baseline factors listed (except ECOG PS; see table 1) encompass entire population (n=312); pairs of n values yielding total <312 are due to unavailability of specific information for all patients.

33

Today's activity

COVID-19 vaccine efficacy trials

- **Primary endpoint**: Symptomatic disease of any severity
- **Time origin**: 7/14 days after last dose
- **Target events**: 150-160
  - 90% power to detect a vaccine with ~60% efficacy.
  - Null hypothesis is that VE ≤ 30%, not VE ≤ 0%.
- **Interim monitoring** strategies vary